

# Statistical aspects of heterogeneous population dynamics

**Kristensen, Kasper; Nielsen, Søren Feodor; Jacobsen, Martin; Lewy, Peter; Thygesen, Uffe Høgsbro**

*Publication date:*
2009

*Document Version*
Publisher's PDF, also known as Version of record

# Statistical aspects of heterogeneous population dynamics

Kasper Kristensen

Thesis submitted for the Ph.D. degree
Institute for Mathematical Sciences
Faculty of Science
University of Copenhagen

Supervisors:

Søren Feodor Nielsen, Institute for Mathematical Sciences
Martin Jacobsen, Institute for Mathematical Sciences
Peter Lewy, Danish Institute for Aquatic Resources
Uffe H. Thygesen, Danish Institute for Aquatic Resources

## Summary

This thesis combines population dynamical models of fish with statistical models of count data obtained from scientific surveys. The aim is to be able to draw conclusions about the biological processes driving the population on basis of observed data. This main problem is addressed using maximum likelihood-based approaches. As a prerequisite it is necessary to give a realistic description of the random variability in the data. The variation is treated as a sum of a contribution due to errors in the population model (system noise) and a contribution caused by errors in the observation process (measurement noise) which occurs because the fish are not distributed uniformly in the sea or because the fish moves.
A length-based population dynamical model is formulated (Section 5.1) and it is shown that the corresponding system-noise is determined by a Poisson process and thus is negligible for large populations. Therefore the primary focus of this thesis is the random variation in the sampling process including the variation due to spatial and time heterogeneity and size-dependent clustering.
The first attempt towards a heterogeneous description of the trawl data is to model the measurement noise using the negative binomial distribution. In combination with the population dynamical model a maximum-likelihood based stock-assessment model is obtained (section 6.3) which allows for formal testing of the underlying biological processes. The statistical model accounts for over-dispersion of the data but offers no opportunity to describe present correlations between size classes.
These correlations may explain apparent changes in catchability from year to year and therefore has major impact on the interpretation of the signal in the data.
This problem is solved by introducing the log Gaussian Cox process which allows for incorporation of correlations in the count data (section 7.1). Correlation structures are formulated to describe spatial heterogeneity of fish on various spatial scales and to deal with the fact that fish of a particular size tend to school with fish of similar sizes.
The application of the log Gaussian Cox process requires special numerical attention (section 7.4) for cases involving large amounts of data.
The log Gaussian Cox process is applied as a substitute of the negative binomial distribution (Section 8.2) in combination with the population model. A further application use the log Gaussian Cox process to estimate concentration areas of fish (section 8.3).
The above considerations are the starting point of the four articles at the end of the thesis.

## Dansk resumé

Denne afhandling beskæftiger sig med at kombinere populationsdynamiske modeller for fisk med statistiske modeller for tælledata opnået fra videnskabelige togter. Formålet er at kunne drage konklusioner om de biologiske processer der driver bestandens udvikling på baggrund af observerede data. For at behandle dette problem gennem maksimum likelihood baserede metoder er det nødvendigt at give en realistisk beskrivelse af den tilfældige variation i data. Denne variation kan naturligt opdeles som et bidrag der skyldes fejl i populations modellen (system-støj) samt et bidrag der skyldes fejl i observations processen (måle-støj) der f.eks. opstår fordi fisk ikke fordeler sig homogent i havet eller fordi fiskene bevæger sig.
En længdebaseret populationsdynamisk model formuleres (sektion 5.1) og der gøres rede for at systemstøjen i denne model er bestemt ved en Poisson proces og dermed er forsvindende for store populationer. Det primære fokus for denne afhandling er derfor den tilfældige variation i sampling processen - herunder variation der skyldes rumlig og tidslig heterogenitet samt størrelses afhængig klumpning.
Det første forsøg i retning af en heterogen beskrivelse af trawl data er at modellere målestøjen ved hjælp af den negative binomial fordeling. I kombination med populationsmodellen fås herved en maksimum likelihood baseret bestandsvurderingsmetode (sektion 6.3) som tillader formel testning af de bagvedliggende biologiske processer. Den statistiske model tager højde for overspredning i data men rummer ikke mulighed for at beskrive tydeligt forkomne korrelationer mellem størrelsesklasser.
Disse korrelationer kan forklare tilsyneladende ændringer i fangbarhed fra år til år og har derfor afgørende indflydelse på fortolkningen af signalet i data. Dette problem løses ved indførelse af den log Gaussiske Cox proces der giver mulighed for at inkorporere korrelationer i tælledata (sektion 7.1). Korrelationsstrukturer formuleres til at tage højde for at fiskene fordeler sig klumpet på forskellige rumlige skalaer samt at fisk af en given størrelse har tendens til at gruppere sig med fisk af samme størrelse i stimer.
Anvendelsen af den log Gaussiske Cox proces kræver særlige numeriske metoder (sektion 7.4) for store mængder af data.
Den log Gaussiske Cox proces anvendes som erstatning for den negative binomial fordeling (sektion 8.2) i kombination med den populationsdynamiske model. Desuden betragtes en anvendelse af den log Gaussiske Cox proces til at estimere koncentrationsområder for fisk (sektion 8.3).
Ovenstående overvejelser danner udgangspunktet for de fire artikler i slutningen af afhandlingen.

# Contents

# Introduction

## 1 Background

Stock assessments are made regularly by fisheries research institutes to aide managers in their regulation of fisheries. Most of the standard assessment models used for this task are estimation algorithms that do not allow for statistical inference (e.g. the XSA model (Shepherd, 1999)).
Standard stock assessment rely heavily on *commercial catch data* which are samples of the fishermens catches. The quality of these data is doubtful due to an increasing amount of fish catches being either non-reported or misreported and a number of assessments are for this reason considered unreliable (ICES., 2005). There is therefore a need to further develop statistical methods that enables stock assessment to be derived from *fishery independent data*, i.e. from scientific bottom trawl surveys.
Stock assessments are typically based on individual age groups where the aging relies on interpretations of ring structures as otoliths or scales. Marine animals that lack such ring structures (e.g. crustaceans) can not be aged this way and for a number of fish stocks poor contrast in the structures impedes reliable aging. For such cases the interpretation of the age structure must be based on the length distribution of the animals. There exist a number of methods that convert length distribution to age (Bhattacharya, 1967; Macdonald and Pitcher, 1979) and it is common practice to use the age-data obtained from these procedures as raw-data in the deterministic XSA-model disregarding the statistical uncertainties.
More recent methods attempts to incorporate length-information in assessment models more rigorously using dynamical models of the length-distributions in conjunction with real statistical models of catch observations (Sullivan, 1992; Frøysa et al., 2002; Schnute and Fournier, 1980; Fournier et al., 1998).
The first component *dynamical length-based population modelling of fish* emerge from the more general ecological discipline of dynamical modelling of structured populations (Metz and Diekmann, 1986). These models are discretized versions of the deterministic flow models based on the von Foerster differential equations (von Foerster, 1959) describing how the size-composition of a population evolves governed by the fundamental biological properties of the individuals of the population recruitment, growth and mortality.

The second component *statistical modelling of catch observations* links expected catches with the observations through standard distributions such as the normal (Sullivan, 1992), the log-normal (Frøysa et al., 2002; Fu and Quinn, 2000) and the multinomial distribution (Schnute and Fournier, 1980; Smith et al., 1998).

## 2 Purpose

The overall purpose of the thesis is to improve the statistical interpretation of trawl-survey data and to demonstrate how the statistical results can be used to extract biological information from the data. The aim is to combine a purely length-based population model with a realistic statistical model of scientific trawl-survey catches.
To this end we find it important to distinguish between *system noise* and *measurement noise*. System noise arise in a population model if stochasticity is added to the biological processes driving the population. Measurement noise reflects the variation of samples conditional on the underlying size distribution of the population.
The key to more realistic statistical description of the measurement noise in fish abundance data is to view the fish-populations as being spatially heterogeneous. We try to give a point-process motivation for the applied distributions even though point-process data are not available. It is a main criterion that the methods have to be computationally feasible in practice with the relatively large amount of data which is available.

## 3 Contributions

**Length-based population modelling** An individual based model of the size distribution of fish is conveniently formulated within a point-process framework. This approach has not been taken elsewhere in the literature. We show that an individual based model including recruitment, mortality and stochastic growth leads to a Poisson process of the entire population (section 5.1) and the intensity is derived. In the special case of deterministic growth the intensity solves the classical deterministic differential equations of von Foerster (1959).

**Statistical interpretation of trawl-survey data** Various statistical distributions have been applied to describe trawl survey data comprising the normal (Sullivan, 1992), the log-normal (Frøysa et al., 2002; Fu and Quinn, 2000) and the multinomial distribution (Schnute and Fournier, 1980; Smith et al., 1998).
The typical large fraction of zeros in trawl survey data has been treated

by extending the log-normal distribution with an atom in zero (Pennington, 1996). Size correlations in trawl survey data have been described by Dirichlet-multinomial and Gaussian-multinomial distributions (Hrafnkelsson and Stefansson, 2004).
Our main contribution is to introduce the log Gaussian Cox process (LGCP) to model spatio-temporal and size correlation in bottom trawl surveys.
We formulate correlation structures to capture relevant heterogeneity.

**Numerical methods for the LGCP** Numerical methods for statistical inference for the LGCP are well-established both in a Bayesian an freqentist setup through MCMC techniques (Møller and Waagepetersen, 2004). These techniques are very general and standard implementations are available e.g. through the R-package (Baddeley and Turner, 2005).
However MCMC-techniques can be very computational expensive. It is well recognized that the simulation based approaches are often outperformed by direct methods such as the Laplace approximation (Skaug and Fournier, 2006) and variants there of (Rue et al., 2007). The approach taken by Skaug and Fournier (2006) uses the Laplace in combination with reverse mode automatic differentiation (Griewank, 2000) to perform approximate ML-estimation. This method is suitable for generalized non-linear mixed models (GNLMMs) containing a moderate number of fixed effects and random effects ($\approx 500 - 1000$). In its direct form the Laplace approximation is unsuitable for GNLMMs with a larger number of random effects because of the need to factorize a second-order derivative matrix of the same dimension as the number of random effects. However, for many interesting models the second-order derivative required by the Laplace approximation contains mostly zeros. Therefore numerical methods for sparse matrices have been considered to make the Laplace approximation feasible for problems involving large data sets (Rue et al., 2007; Rue, 2005; Rue et al., 2004; Bates, 2004). The approach of Bates (2004) implemented in the R-package "lme4" (Bates et al., 2008) handles GLMMs but is limited to covariance structures which can be expressed through a (well-designed) formula interface.
Our contribution mixes ideas of the existing numerical methods in order to handle the LGCP in cases with large amounts of data and non-linear geostatistical covariance structures. Inspired by Rue and Held (2005) we restrict attention to covariance structures with a sparse inverse - the so-called Gaussian Markov Random fields (GMRFs). Like Bates (2004) our approach uses an augmented system to take full advantage of sparseness and to gain numerical stability.
We finally develop a quadratic approximation of the LGCP-likelihood which is cheap to evaluate in practice. The quadratic approximation is used for fitting and testing non-linear models of the fixed effects of the LGCP.

# 4 Paper overview

**Paper I** The simplest step towards an underlying heterogeneous interpretation of the statistical distribution of fish is to apply a distribution which allows for over-dispersion. Can a negative binomial distribution adequately describe observed size-distribution if it is combined with a length-based model of a fish stock? This is examined in paper I. The main conclusion is that it is possible to carry out a length-based stock assessment based on relatively few survey observations even with the high degree of over-dispersion in the data. However, over-dispersion is not the only problem with the data. High correlations between the number of fish in neighboring size-classes are encountered which the negative binomial distribution does not account for. These issues are the main focus of the following three papers.

**Paper II** To deal with the correlations the LGCP is considered. It has previously been used to describe heterogeneity of e.g. animals and plants in numerous ecological studies. It is also suitable for statistical modelling of length-based trawl-survey data because of its ability to model high-dimensional correlated count data. A correlation structure is formulated in order to capture the random effect of a large-scale spatio temporal log-abundance surface and small-scale size dependent clustering. An ML-estimation algorithm based on the Laplace approximation is formulated. The method is aimed at large sparse precision matrices for which modern sparse matrix solvers can be used to make the estimation practically possible. It is shown how the specified correlation structure can be given a formulation for which the precision matrix is sparse. The method is applied on a single survey in the North-Sea.

**Paper III** The length-based model from "paper I" is combined with the size-space-time-correlated LGCP from "paper II" in order to fix the lacking correlations in the negative binomial distribution. The main question we wish to answer is whether there are remarkable changes in the conclusion about the biological length-based population model when data is interpreted through the more realistic LGCP. It is concluded that the inclusion of size-space and time correlations generally increase the precision of the size-spectrum slope while the precision of the overall spectrum level is decreased. As an important consequence a time-changing catchability is not significant as opposed to the conclusions of "paper I".

**Paper IV** A statistical model which accounts for spatial correlation is suitable for spatial prediction. It is thus obvious to use the LGCP for spatial interpolation of fish-abundance surfaces. A prediction method based on a statistical model is convenient because the statistical model can be validated

as opposed to existing ad hoc methods.
Data of North-Sea cod is considered and the LGCP is fitted with a three-parameter spatial correlation structure separately for each of three age groups during the period 1983-2006. The model is accepted using residual-based goodness of fit assessment.
Time-changes in various concentration measures are examined. In particular $D_{95}$ - the smallest fraction of the area containing 95% of the population - is considered as a function of the hidden intensity. It is concluded that the posterior mean of $D_{95}$ given the data is unchanged during the period. This observation contradicts the theory of the ideal free distribution.

# 5 Length-based population modelling

## 5.1 Individual based formulation

Size based population models attempts to create the link between biological knowledge about the single individual and the size distribution of an entire population assuming that individuals share some fundamental biological properties. Theses issues are known as *scaling problems* within the biological field.
It is commonly recognized that the size-distribution of a fish-population is mainly governed by the fundamental biological processes recruitment, growth and mortality.
As an example of an individual based biological model of recruitment, growth and mortality consider the following individual assumptions:

1. An individual is born (recruited) during the small time interval $[t, t + \Delta t]$ with probability $r(t)\Delta t + o(\Delta t)$ independent of the past where $r$ is the recruitment function.

2. An individual of size $x$ dies during the small time interval $[t, t + \Delta t]$ with probability $z(x, t)\Delta t + o(\Delta t)$ independent of the past where $z$ is the size- and time specific mortality rate.

3. An individual born at time $s$ grow according to a stochastic growth trajectory $L_s(t)$.

4. Individuals grow and die independently.

The individual model is conveniently visualized (Fig. 1a) by representing each individual with its growth-curve. The recruitment process then appears as points on the time-axis while the size-distribution of individuals alive at time $t$ appears as crossings of the growth-curves with a vertical (dashed) line. An interrupted growth-curve indicates the death of an individual.

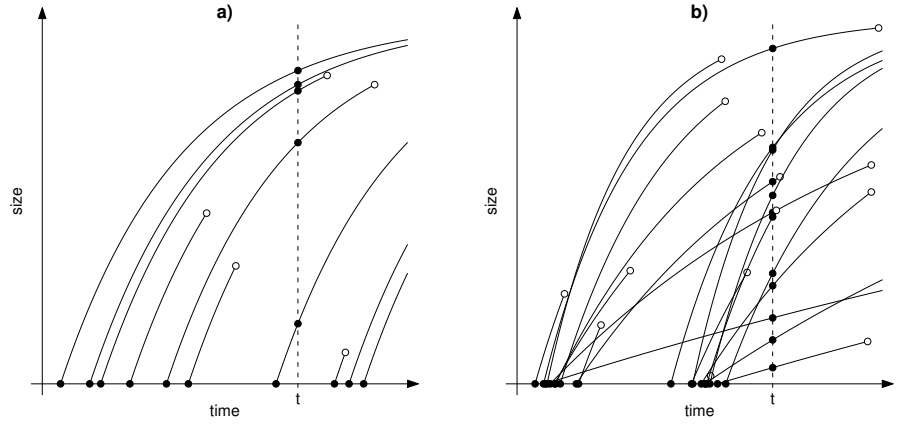



**Figure 1:** Illustration of individual based model of growth, mortality and recruitment. A solid circle on the time axis indicates a recruitment event. An open circle indicates the death of an individual. Crossings of the vertical dashed line with the growth trajectories are marked with a solid circle to indicate the individuals alive at time $t$. (a) Constant recruitment-rate and deterministic growth-curves. (b) Time-inhomogeneous recruitment and stochastic growth-curves.

The first assumption characterizes the point-process on the time-axis as being an inhomogeneous Poisson-process with intensity $r$. To solve the scaling problem we must find the distribution of the "vertical" point-process which keeps track of the size-distribution of the individuals alive at time $t$. Denote by $N_t$ this counting process defined by letting $N_t(A)$ be equal to the number of individuals alive at time $t$ with a size contained in $A \subset R$. To find its distribution note that the remaining assumptions (2-4) suggests an *independent random labeling* (Møller and Waagepetersen, 2004) of the recruitment-process. Indeed for any configuration of disjoint sets $A_1, ..., A_k \subset R$ attach the label $A_i$ to a recruitment-point $s$ if the corresponding individual is alive at time $t$ with a size contained in $A_i$ - and denote by $P_{s,t}(A_i)$ the probability of this event. It follows that the recruitment process split across labels constitutes independent Poisson-processes with intensities $s \to r(s)P_{s,t}(A_i)$ (Møller and Waagepetersen, 2004). In turn the random variables $N_t(A_1), ..., N_t(A_k)$ becomes independent Poisson distributed with mean $E(N_t(A_i)) = \int_0^t r(s)P_{s,t}(A_i)\, ds$. In conclusion $N_t$ is again an inhomogeneous Poisson-process with intensity $\lambda(x,t) = \frac{\partial}{\partial x}\int_0^t r(s)P_{s,t}([0,x])\, ds$.
Next consider the probability $P_{s,t}([0,x])$ that an individual born at time $s$ is still alive at time $t$ with a size included in the set $[0,x]$. According to assumption 2 the hazard function of an individual following a fixed growth-trajectory $l_s$ initiated at time $s$ is $\tau \to z(l_s(\tau),\tau)$. Thus the probability of survival up to

time $t$ is $\exp\left(-\int_s^t z(l_s(\tau),\tau)\,d\tau\right)$. To find $P_{s,t}([0,x])$ for a general stochastic growth curve $L_s(\tau)$ initiated at time $s$ we take expectation over the possible growth-curves $P_{s,t}([0,x]) = E\left(\exp\left(-\int_s^t z(L_s(\tau),\tau)\,d\tau\right)1_{(L_s(t)\leq x)}\right)$. Insert this to get the general expression of the intensity of $N_t$

$$\lambda_t(x) = \frac{\partial}{\partial x}\int_{-\infty}^{t} r(s)E\left(\exp\left(-\int_s^t z(L_s(\tau),\tau)\,d\tau\right)1_{(L_s(t)\leq x)}\right)\,ds \qquad (1)$$

The intensity (1) completely specifies the distribution of the population size composition. The individual based model includes the effect of stochastic recruitment, mortality and growth (Fig 1b). It may therefore appear somewhat surprising that this biological system creates no more than Poisson variation in the output-process $N_t$. For a large population the Poisson noise hardly matters and it is tempting to think of the population size-distribution as a deterministic process.

### 5.2 Stochastic von Bertallanfy growth

The particular form of the growth model applied in thesis takes its starting point in the classical von Bertallanfy growth model (Bertalanffy, 1938):

$$L_s(t|L_\infty,k,L_0) = L_\infty - (L_\infty - L_0)e^{-k(t-s)} \qquad (2)$$

This equation describes the growth of an individual born at time $s$. The growth trajectory approach the asymptotic size $L_\infty$ as $t$ tends to infinity.
A stochastic growth-model is obtained by assuming that each individual is assigned its personal asymptotic size $L_\infty$ chosen from a common distribution with density $u$ on $[L_0,\infty)$. All individuals are assumed to have the same growth parameter $k$.
To find the intensity (1) in this case note first that at time $t$ the individuals that have size less than $x$ are exactly the ones with an $L_\infty$ belonging to the set

$$\{L_\infty : L(t,L_\infty) \leq x\} = [L_0, G(x)] \qquad (3)$$

where

$$G(x) = G(x|k,s,L_0) = \frac{x - L_0e^{-k(t-s)}}{1-e^{-k(t-s)}} \qquad (4)$$

Now equation (1) becomes

$$\begin{aligned}\lambda_t(x) &= \frac{\partial}{\partial x}\int_{-\infty}^{t} r(s)E\left(\exp\left(-\int_s^t z(L_s(\tau|L_\infty),\tau)\,d\tau\right)1_{(L_s(t)\leq x)}\right)\,ds \\ &= ... \\ &= \int_{-\infty}^{t} r(s)\exp\left(-\int_s^t z(L_s(s,G_s(x)),s)\,ds\right)u(G_s(x))G_s'(x)\,ds\end{aligned} \qquad (5)$$

### 5.3 Traditional formulation

Traditional modelling of population size-distributions takes its starting point in the von Foerster PDE (von Foerster, 1959)

$$\frac{\partial}{\partial t}n(x,t) = -\frac{\partial}{\partial x}(g(x,t)n(x,t)) - z(x,t)n(x,t) \tag{6}$$

with the boundary condition $r(t) = n(0,t)g(0,t)$ and $g(x,t)$ denotes the growth-rate of an individual of size $x$ at time $t$. In this context $n(x,t)$ is called the "number-density" and has the property that $\int_A n(x,t)\,dx$ is the deterministic number of individuals with size contained in $A$ at time $t$.
It is straight forward to show that in the case of deterministic growth the intensity (1) solves (6) and thus the concept of "intensity" and "number-density" are identical.
The stochastic growth model from the previous section could alternatively be obtained by treating solutions to the von-Foerster equation as function of $L_\infty$ and the mixing all these solutions wrt. the probability density $u$. However, this approach is very inappropriate from a numerical perspective. The more direct form (5) is easier to handle in practice. A discretization of the inner integral is known as the method of *integration along characteristics* and is a recognized way to solve the differential equations efficiently.

## 6 The inverse problem

The main issue of interest is to estimate the biological processes recruitment, mortality and growth based on samples of individual sizes. Having Fig. 1 in mind what can we say about the biological system based on samples of the vertical point-process? This inverse problem can be formulated within a maximum likelihood framework if we can specify how the available samples are collected from the population.

### 6.1 Data

The data considered in this thesis are obtained from scientific bottom trawl surveys. The survey is conducted by vessels following a randomized route covering the population area of interest. At each of the chosen positions a sample (haul) is taken with the trawl. The duration and speed of the trawl is approximately the same for all samples and thus a sample is prescribed to cover a given *swept area*.
As the spatial positions of the trawl is random any fish must have the same probability of belonging to the swept area at the time of the sample. Denote by $p$ this probability given as the ratio of swept area and total population area. Whether a fish within the swept area is caught obviously depends on the fish size. A small fish will have a higher chance of escaping through

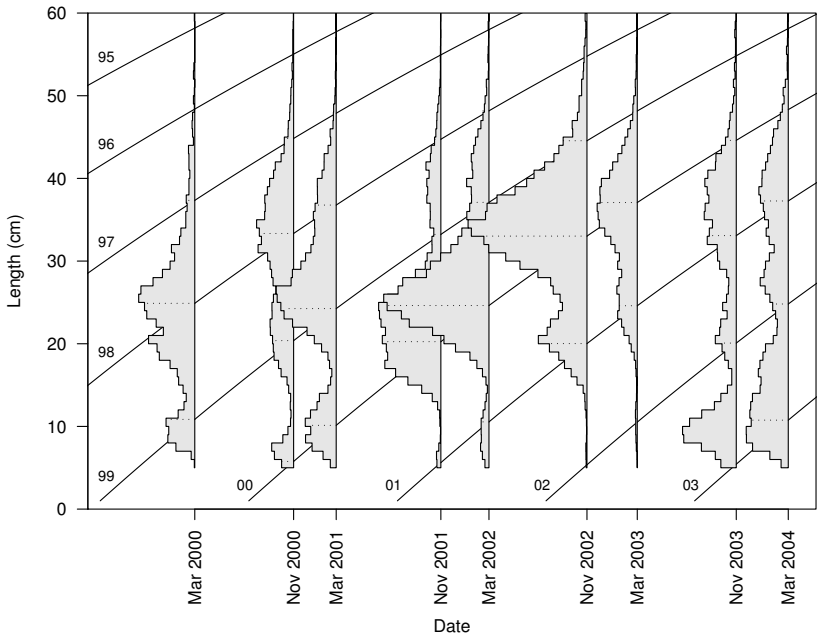



**Figure 2:** Average CPUE as function of size for each of nine surveys of Baltic-cod. Lines indicate von Bertallanfy curves with parameter values from Bagge et al. (1994).

the meshes than a larger fish. This phenomenon is known as *gear selectivity* and can be modelled by a selectivity function $s(x)$ denoting the conditional probability that a fish of size $x$ gets caught given its presence within the swept-area.
Note that despite the concrete interpretation of $p$ its value is unknown because we do not know the extension area of the population. For the same reason $p$ could be time-dependent.

## 6.2 The Poisson model

A first (naive) attempt to formulate a statistical model of the samples is to argue that any fish of size $x$ from the population has probability $ps(x)$ of ending up in a given sample. Thus a haul can be viewed as an *independent random thinning* (Møller and Waagepetersen, 2004) of the population with thinning probability $ps(x)$. A sample is then a realization of a Poisson-process with intensity

$$\lambda_t^{obs}(x) = ps(x)\lambda_t(x) \tag{7}$$

because the population was described through a Poisson-process with intensity (1). The expected number of fish $N_C$ in a length-class $C$ is then

$$E(N_C) = \int_C ps(x)\lambda_t(x)\,dx \tag{8}$$

Based on these expected values we can in principle write down the corresponding Poisson-likelihood and for given parameterizations of the biological processes carry out maximum-likelihood estimation.

### 6.3 The negative binomial model

One of the first practical things to learn about trawl-survey data is that they are almost all very far from being Poisson distributed. The first attempt to solve this problem is to replace the Poisson distribution with a distribution allowing for over-dispersion. This is done in *Paper I* (page 35) which we briefly describe in the following.
Let $N_{ij}$ denote the observation matrix of counts of $i$th haul and $j$th length-group. Associate with $i$ the corresponding survey $survey_i$. As the hauls within a given survey are taken within a relatively short time-interval it is reasonable to assume that the size distribution of the fish-population is unchanged during the survey. Thus our main model states that

$$E[N_{i,j}] = \mu_{survey_i,j} \tag{9}$$

where the parameter matrix of $\mu_{t,j}$ holds the size-composition of survey $t$. We do not impose any restrictions on the variance of the counts and associate with each mean-value parameter a free variance parameter

$$V[N_{i,j}] = \sigma^2_{survey_i,j} \tag{10}$$

Assuming the counts follows a negative binomial distribution and that all counts are independent the likelihood is

$$L((\mu_{t,j}),(\sigma^2_{t,j})) = \prod_i \prod_j \frac{\Gamma(N_{ij}+\nu_{t_i,j})}{\Gamma(\nu_{t_i,j})\Gamma(N_{ij}+1)} \pi_{t_i,j}^{\nu_{t_i,j}} (1-\pi_{t_i,j})^{N_{ij}} \tag{11}$$

where $\pi_{t_i,j} = \frac{\mu_{t_i,j}}{\sigma^2_{t_i,j}}$ and size parameter $\nu_{t_i,j} = \frac{\mu^2_{t_i,j}}{\sigma^2_{t_i,j}-\mu_{t_i,j}}$. To reduce the number of parameters in the main model we state the variance structure hypothesis

$$\sigma^2_{t,j} = a_t \mu^{b_t}_{t,j} + \mu_{t,j} \tag{12}$$

This a more flexible structure than the common assumption of a fixed $\nu$-parameter across groups corresponding to the special case of $b_t = 2$ in (12). The variance structure (12) is a submodel of the un-restricted variance model (10) and can thus be formally tested with a likelihood-ratio test.
Likewise the length-based population model (1) can be treated as a sub-model of the general un-restricted mean-value model (9):

$$\mu_{tj} = \int_{C_j} ps_\theta(x)\lambda_\theta(x,t)\,dx \tag{13}$$

where $C_j$ represents the $j$th size-interval. Both the gear-selectivity and intensity now depends on an unknown parameter vector $\theta$ which is to be estimated. The chosen parametric form of the biological processes is

1. The recrutiment $r_\theta(t)$ is a linear combination of yearly varying Gaussian peaks (3 parameters per year).

2. The distribution of $L_\infty$ is chosen to be normal (2 parameters).

3. The size-specific mortality is a sum of a constant natural mortality and sigmoid size-dependent fishing mortality with a yearly varying asymptotic level (3 parameters plus one parameter per year).

4. Survey selectivity $s_\theta(x)$ is chosen as a sigmoid function of fish-size (2 parameters).

For more details about the parameterization we refer to *Paper I*.
Insertion of (13) and (12) in the likelihood (11) yields the likelihood under the hypothesis of the corresponding size-structured population model. It is not obvious whether parameter-estimation is possible in this model. First thing to notice is that if $p$ and $r_\theta(t)$ are multiplied and divided respectively with the same constant then the likelihood is unchanged. This fact just reflects that it is only possible to estimate the recruitment relatively. A solution is to fix the recruitment for one of the years.
It is shown in *Paper I* by extensive simulation studies that it is possible to re-estimate known parameters from simulated data-sets and that standard asymptotic likelihood theory applies for this estimation problem. The method is applied on a collection of nine surveys in the Baltic (Fig. 2). Based on this relatively small data-set it is possible to estimate the parameters even with the substantial level of over-dispersion in the data.
It is an important strength of the likelihood approach that it permits formal testing of the validity of the length-based population model and sub-models. For instance it is relevant from a management point of view to be able to judge whether there is a significant change in fishing-mortality from one year to the next. However, formal testing requires a valid statistical model. While the negative binomial distribution describes the marginals nicely it is pointed out that there are clear signs of correlations in the data which are not accounted for. Empirical correlations between neighboring length-classes within the same survey are higher than 90% and the correlation range appears to span more that 15-20 cm (Fig. 3).
The tests must be considered as unreliable as the model ignores the correlations.
These issues are considered in *Paper II* and *Paper III*.

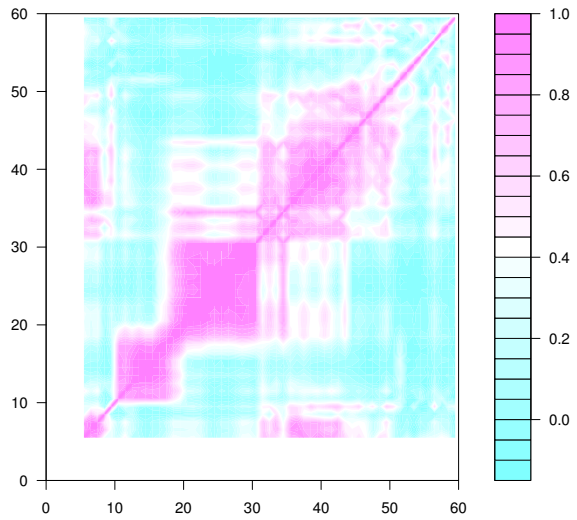



**Figure 3:** Image of empirical correlation matrix of the autumn 2001 survey of cod in the Baltic in which a total of 8610 fish were caught in 33 hauls.

# 7 Incorporating correlations in the observation model

Clearly there is a need to introduce correlations in the statistical distribution of trawl-survey data. Instead of just choosing an arbitrary distribution we find it convenient to seek inspiration in existing point-process models because our sampling problem has a natural point-process interpretation. A fish population may be thought of as a heterogeneous spatial point pattern changing dynamically in time. Each point is given an "attribute" in terms of the fish size (Fig. 4). Fish samples taken with a trawl can be thought of as a size-dependent random thinning of the point pattern within a rectangular region.
We restrict attention to the so-called Cox-processes

## 7.1 Log Gaussian Cox-process

The log-Gaussian-cox process (LGCP) is a Cox-process with random log-intensity following a Gaussian process (Møller et al., 1998). We give a formulation suitable for spatio-temporal modeling of the size-composition of fish. Let $\eta(s,x,t)$ denote a Gaussian random field indexed by size, space and time respectively. For any point in time $t$ let $N_t$ be a Poisson-process with intensity $(\exp(\eta(s,x,t)))_{(x,s)\in R^2\times R_+}$. Then for any haul-rectangle $H \subset R^2$ and size-class $C \subset R_+$ the conditional distribution of the number of points

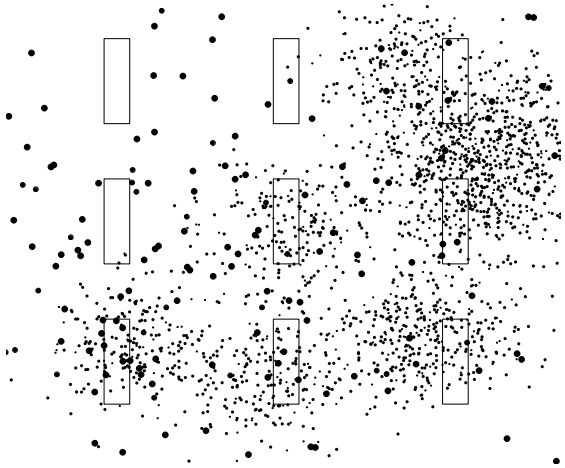


**Figure 4:** Illustration of size-dependent clustering. Fictive positions of individual fish in a point-process setup where the points are marked with the individual sizes (only two sizes are considered for simplicity) and nine fictive haul-rectangles.

in a $H \times C$ given $\eta$ is

$$N_t(H \times C)|\eta \sim Pois\left(\int_H \int_C e^{\eta(s,x,t)}\, ds\, dx\right) \tag{14}$$

This equation specifies the distribution of the number of points within a rectangle for the various size-classes (Fig. 4). Size-selectivity was previously introduced as the conditional probability that a fish is caught given its presence within the haul-rectangle. Denote by $q(s)$ this probability. After a random thinning the observed number of points within the rectangle is (Møller and Waagepetersen, 2004)

$$N_t^{obs}(H \times C)|\eta \sim Pois\left(\int_H \int_C q(s) e^{\eta(s,x,t)}\, ds\, dx\right) \tag{15}$$

From a large-scale perspective it is reasonable to assume the intensity is approximately constant within the haul-rectangle leading to the approximation

$$N_t^{obs}(H \times C)|\eta \sim Pois\left(q(s) e^{\eta(s,x,t)} |H||C|\right) \tag{16}$$

for some $(x, s) \in H \times C$. This distribution is just a multivariate Poisson distribution with a multivariate log-normal intensity.

## 7.2 LGCP likelihood

Likelihood inference for the model along with the computational issues will be discussed in the following.

Let

$$\eta \sim N(\mu, \Sigma_\theta)$$
$$N|\eta \sim \otimes_{i=1}^{n} Pois(\eta_i)$$

The full negative log-likelihood where both $\eta$ and $N$ are observed is given by

$$l_{full}(\theta, \mu|\eta, N) = \sum_{i=1}^{n} e^{\eta_i} - \sum_{i=1}^{n} N_i\eta_i - \frac{1}{2}\log\det Q_\theta + \frac{1}{2}(\eta-\mu)^t Q_\theta(\eta-\mu) + c$$

where $Q_\theta = \Sigma_\theta^{-1}$ is the precision and $c = \frac{n}{2}\log(2\pi) + \sum_{i=1}^{n}\log\Gamma(N_i+1)$.
The marginal likelihood - for unobserved $\eta$ - is

$$l(\theta, \mu|N) = -\log\left(\int_{R^n} \exp(-l_{full}(\theta, \mu|\eta, N))\, d\eta\right) \tag{17}$$

The integral is difficult to evaluate numerically. In the following we go through a standard method - the Laplace approximation - for approximating high dimensional integrals based on a Gaussian approximation of the conditional distribution of $\eta|N$.

### 7.3 Laplace approximation

Several authors have good experience with the Laplace approximation because its level of accuracy is often high compared to the computational cost (Rue et al., 2007; Skaug and Fournier, 2006). The Laplace approximation has become the standard method for fitting GLMMs in R (Bates et al., 2008).
In the following a brief description of the Laplace approximation is given. With starting point in (17) consider the problem of approximating an integral of the form

$$-\log\int \exp(-f(\theta,\eta))\, d\eta$$

i.e. the negative log-likelihood of a mixed model with random parameters $\eta$ where $f(\theta,\eta) = l_{full}(\theta|x,\eta)$ is the negative log-likelihood of the full model where the random parameters are observed.
Let $\hat{\eta}_\theta$ be the argument of minimum of $f$ for fixed $\theta$

$$\forall\theta \quad f'_\eta(\theta, \hat{\eta}_\theta) = 0 \tag{18}$$

A Taylor-expansion gives:

$$f(\theta,\eta) \approx f(\theta,\hat{\eta}_\theta) + \frac{1}{2}(\eta-\hat{\eta}_\theta)^t f''_{\eta\eta}(\theta,\hat{\eta}_\theta)(\eta-\hat{\eta}_\theta)$$

and the integral may be approximated by

$$\int \exp(-f(\theta,\eta))\, d\eta \approx \exp(-f(\theta,\hat{\eta}_\theta)) \int \exp\left(-\frac{1}{2}(\eta-\hat{\eta}_\theta)^t f''_{\eta\eta}(\theta,\hat{\eta}_\theta)(\eta-\hat{\eta}_\theta)\right) d\eta$$
$$= \exp(-f(\theta,\hat{\eta}_\theta))\frac{(2\pi)^{\frac{n}{2}}}{\sqrt{\det f''_{\eta\eta}(\theta,\hat{\eta}_\theta)}}$$

where $n$ is the dimension of the random parameter space. Hence we have the negative log marginal likelihood approximated by.

$$-\log \int \exp(-f(\theta,\eta))\, d\eta \approx f(\theta,\hat{\eta}_\theta) - \frac{n}{2}\log 2\pi + \frac{1}{2}\log\det f''_{\eta\eta}(\theta,\hat{\eta}_\theta) \quad (19)$$

The gradient can be useful for efficient optimization of (19). Taking derivative of (18) wrt. $\theta$ gives:

$$f''_{\eta\theta}(\theta,\hat{\eta}_\theta) + f''_{\eta\eta}(\theta,\hat{\eta}_\theta)\frac{d}{d\theta}\hat{\eta}_\theta = 0 \implies \frac{d}{d\theta}\hat{\eta}_\theta = -f''_{\eta\eta}(\theta,\hat{\eta}_\theta)^{-1} f''_{\eta\theta}(\theta,\hat{\eta}_\theta) \quad (20)$$

Define

$$h(\theta,\eta) = f(\theta,\eta) - \frac{n}{2}\log 2\pi + \frac{1}{2}\log\det f''_{\eta\eta}(\theta,\eta)$$

Then the desired gradient is given by

$$\frac{d}{d\theta}h(\theta,\hat{\eta}_\theta) = h'_\theta(\theta,\hat{\eta}_\theta) - h'_\eta(\theta,\hat{\eta}_\theta) f''_{\eta\eta}(\theta,\hat{\eta}_\theta)^{-1} f''_{\eta\theta}(\theta,\hat{\eta}_\theta) \quad (21)$$

This formula is also stated in Skaug and Fournier (2006).
Returning now to the case of the LGCP likelihood the formulas for computing the Laplace approximation and its gradient are:

$$f'_\eta(\theta,\eta) = e^\eta - N + Q_\theta(\eta-\mu)$$

$$f'_\mu(\theta,\eta) = -Q_\theta(\eta-\mu)$$

$$f'_{\theta_i}(\theta,\eta) = -\frac{1}{2}tr(Q_\theta^{-1}\dot{Q}_\theta) + \frac{1}{2}(\eta-\mu)^t\dot{Q}_\theta(\eta-\mu)$$

$$f''_{\eta\eta}(\theta,\eta) = diag(e^\eta) + Q_\theta$$

This 2nd order derivative is everywhere positive definite which implies strictly convexity. So the inner likelihood has a unique minimum.

$$f''_{\eta\mu}(\theta,\eta) = -Q_\theta$$

$$f''_{\eta\theta_v}(\theta,\eta) = \dot{Q}_\theta(\eta-\mu)$$

The $h$-function:

$$h(\theta,\eta) = f(\theta,\eta) + \frac{1}{2}\log\det\left(diag(e^\eta) + Q_\theta\right)$$

has derivatives:

$$h'_\eta(\theta,\eta) = f'_\eta(\theta,\eta) + \frac{1}{2}[e^{\eta_i}(f''_{\eta\eta}(\theta,\eta)^{-1})_{ii}]$$

$$h'_\mu(\theta,\eta) = f'_\mu(\theta,\eta)$$

$$h'_{\theta_i}(\theta,\eta) = f'_\theta(\theta,\eta) + [\frac{1}{2}tr(\dot{Q}_\theta\,(diag(e^\eta)+Q_\theta)^{-1})]$$

These expressions are what we need to compute (21).
Some computational remarks are worth noticing when dealing with the above formulas in practice. The computational complexity can be reduced a lot if $Q_\theta$ is assumed to be *sparse*. For instance consider the computational complexity of $tr(\dot{Q}_\theta\,(diag(e^\eta)+Q_\theta)^{-1})$. The trace of matrix product is the sum of the pointwise product of the matrices so the inverse $(diag(e^\eta)+Q_\theta)^{-1}$ is only needed on the non-zero pattern of $\dot{Q}_\theta$ (which is smaller than or equal to the pattern of $Q_\theta$). An existing algorithm known as the *inverse-subset algorithm* is designed to handle this problem (Rue, 2005).
To perform estimation of the fixed effects in practice we have good experience with the following approach:

- Handle the outer non-linear optimization problem of the fixed effects $(\theta,\mu)$ by the BFGS-method.
- Perform the inner convex optimization problem with an ordinary Newton method.

The Newton method is of course only recommended because the second-order derivative $Q_\theta + diag(e^\eta)$ of the inner likelihood wrt. $\eta$ is assumed sparse.

### 7.4 An augmented system

The special case of a linear mean-value structure $\mu = A\beta$ for a full rank design matrix $A$ is sometimes referred to as a generalized linear geostatistical model (GLGMs) (Diggle and Ribeiro, 2006). The LGCP-likelihood is

$$\begin{aligned} l_{full}(\theta,\beta|\eta,x) = \sum_{i=1}^n e^{\eta_i} - \sum_{i=1}^n x_i\eta_i - \frac{1}{2}\log\det Q_\theta \\ + \frac{1}{2}(\eta - A\beta)^t Q_\theta(\eta - A\beta) + c \end{aligned}$$

with marginal likelihood

$$l(\theta,\beta|x) = -\log\int e^{-l(\theta,\beta|\eta,x)}\,d\eta \tag{22}$$

We shall now see that for this special linear model the fixed effect $\beta$ can be moved from the *outer optimization* to the *inner optimization*.
The exact score of (22) wrt $\beta$ is

$$\nabla_\beta l(\beta,\theta) = -A^t Q_\theta(E_{(\beta,\theta)}[\eta|x] - A\beta) \tag{23}$$

The Gaussian posterior approximation suggests replacing $E_{\beta,\theta}[\eta|x]$ by $\hat{\eta}(x)$. Thus $(\hat{\eta},\hat{\beta})$ can be found simultaneously by solving

$$e^\eta - x + Q_\theta(\eta - A\beta) = 0 \tag{24}$$

$$A^t Q_\theta(\eta - A\beta) = 0 \tag{25}$$

through the corresponding Newton iterations

$$\begin{aligned}\begin{pmatrix}\eta_{k+1}\\ \beta_{k+1}\end{pmatrix} &= \begin{pmatrix}\eta_k\\ \beta_k\end{pmatrix}\\ &- \begin{pmatrix}Q_\theta + diag(e^{\eta_k}) & -Q_\theta A\\ -A^t Q_\theta & A^t Q_\theta A\end{pmatrix}^{-1}\begin{pmatrix}e^{\eta_k} - x + Q_\theta(\eta_k - A\beta)\\ -A^t Q_\theta(\eta_k - A\beta)\end{pmatrix}\end{aligned} \tag{26}$$

This approach has the interpretation of treating the augmented vector $(\eta, \beta)$ as a random effect with (improper) precision

$$\begin{pmatrix}Q_\theta & -Q_\theta A\\ -A^t Q_\theta & A^t Q_\theta A\end{pmatrix} \tag{27}$$

corresponding to the hierarchical model where $\beta$ is drawn from a diffuse prior and $\eta|\beta \sim N(A\beta, Q_\theta^{-1})$.
Even though (27) is only positive semi-definite it is easy to show (using that $A$ has full rank) that the matrix

$$\begin{pmatrix}Q_\theta + diag(e^{\eta}) & -Q_\theta A\\ -A^t Q_\theta & A^t Q_\theta A\end{pmatrix} \tag{28}$$

is positive definite for any $\eta$. This means that in practice the Newton iterations (26) defines a stable optimization problem.
Another important remark is that (28) inherits the sparseness of $Q_\theta$ and $A$ allowing the Newton iterations (26) to be carried out efficiently for large problems.
But is it really necessary to consider an augmented system? - why not just substitute the solution of (25) wrt. $\beta$ into (24) and then solving the reduced system which only involves $\eta$? The answer to this question is that the reduced system is no longer sparse and thus considering the augmented system really is a good idea for computational reasons.
To summarize the above procedure - referred to as the *inner optimization problem* - we have found the posterior mode $\hat{\eta}_\theta$ and ML-estimate $\hat{\beta}_\theta$ jointly for any given $\theta$. By inserting $\hat{\eta}_\theta$ and $\hat{\beta}_\theta$ in the Laplace approximation (19) of (22) we thus obtain an approximate likelihood profile wrt. $\theta$

$$l_{prof}(\theta|x) \approx l_{full}(\theta, \hat{\beta}_\theta|\hat{\eta}_\theta, x) + \frac{1}{2}\log\det\left(Q_\theta + diag(\hat{\lambda}_\theta(x))\right) \tag{29}$$

Optimization of this profile wrt. $\theta$ - the *outer optimization* - is suitable for the BFGS algorithm (Fletcher, 1970) because the objective function is non-linear in $\theta$ and because $\theta$ usually is a relatively short vector. Standard implementations of the BFGS (e.g. "optim" (R Development Core Team, 2008)) finds the Hessian $\nabla^2 l_{prof}(\theta|x)$ as a by-product of the optimization. This Hessian is the approximate precision of $\hat{\theta}$. However, we actually need the joint precision of the entire fixed effect vector $(\hat{\beta}, \hat{\theta})$. Denote by

$$\begin{pmatrix}H_{\beta\beta} & \\ H_{\theta\beta} & H_{\theta\theta}\end{pmatrix} \tag{30}$$

this matrix. The block-matrix $H_{\theta\theta}$ can be found by noting that the Hessian of the profile likelihood defines the marginal precision of (30) (Pawitan, 2001)

$$H_{prof} = H_{\theta\theta} - H_{\theta\beta}H_{\beta\beta}^{-1}H_{\theta\beta}^t$$

so that the full precision (30) becomes

$$\begin{pmatrix} H_{\beta\beta} & \\ H_{\theta\beta} & H_{prof} + H_{\theta\beta}H_{\beta\beta}^{-1}H_{\theta\beta}^t \end{pmatrix} \tag{31}$$

The first block-column is found directly from (23) by differentiation wrt. $\beta$ and $\theta$ respectively.

$$H_{\beta\beta} = \nabla_\beta^2 l(\beta, \theta) \tag{32}$$
$$H_{\theta\beta} = \nabla_\theta \nabla_\beta l(\beta, \theta) \tag{33}$$

We prefer a further rewriting of (31). Recall that the definition of $\hat{\beta}_\theta$ is given implicitly through the equation (23) with the conditional mean replaced by the posterior mode:

$$-A^t Q_\theta(\hat{\eta}_{(\beta,\theta)} - A\hat{\beta}_\theta) = 0 \tag{34}$$

A chain-rule argument similar to (20) then gives the identity

$$\nabla_\theta \hat{\beta}_\theta = -H_{\beta\beta}^{-1}H_{\beta\theta}$$

which suggests rewriting (31) as

$$\begin{pmatrix} H_{\beta\beta} & \\ -G^t H_{\beta\beta} & H_{prof} + G^t H_{\beta\beta} G \end{pmatrix} \tag{35}$$

where $G := \nabla_\theta \hat{\beta}_\theta$. The expressions required to compute (35) are given by

$$H_{\beta\beta} = A^t Q A - A^t Q(Q + diag(e^{\hat{\eta}}))^{-1} Q A$$

and - using the same chain-rule argument on $(\hat{\eta}, \hat{\beta})$

$$\nabla_\theta \begin{pmatrix} \hat{\eta} \\ \hat{\beta} \end{pmatrix} = -\begin{pmatrix} Q_\theta + D_{\hat{\eta}} & -Q_\theta A \\ -A^t Q_\theta & A^t Q_\theta A \end{pmatrix}^{-1} \begin{pmatrix} \dot{Q}_\theta(\hat{\eta} - A\hat{\beta}) \\ -A^t \dot{Q}_\theta(\hat{\eta} - A\hat{\beta}) \end{pmatrix} \tag{36}$$

Lets illustrate the usefulness of formula (35) in practice. For the cases considered in this thesis the dimension of $\beta$ ranges from 60 to 500 while $\theta$ has dimension 6. For these applications the only time-consuming part of computing (35) is to calculate the small 6 by 6 matrix $H_{prof}$. The rest of the calculations takes less than the time of a single likelihood evaluation.
Besides allowing for construction of confidence regions around $(\hat{\beta}, \hat{\theta})$ formula

(35) can be used to obtain a quadratic approximation of the LGCP-likelihood (22) in a neighborhood around $(\hat{\beta}, \hat{\theta})$:

$$l(\theta, \beta|x) - l(\hat{\theta}, \hat{\beta}|x) \approx \frac{1}{2} \begin{pmatrix} \beta - \hat{\beta} \\ \theta - \hat{\theta} \end{pmatrix}^t \begin{pmatrix} H_{\beta\beta} & -H_{\beta\beta}G \\ -G^t H_{\beta\beta} & H_{prof} + G^t H_{\beta\beta} G \end{pmatrix} \begin{pmatrix} \beta - \hat{\beta} \\ \theta - \hat{\theta} \end{pmatrix} \tag{37}$$

Denote by $q(\beta, \theta)$ the right hand side of this display. Relying on standard asymptotic theory one would expect the approximation being accurate within a confidence-region of the form

$$C = \{(\beta, \theta) : 2q(\beta, \theta) < F^{-1}_{\chi^2(n)}(95\%)\}$$

where $n$ is the dimension of the vector $(\beta, \theta)$. Thus the quadratic approximation can be used to fit and test sub-models *independent of numerical integration* required by the true LGCP-likelihood (17).
Consider for instance a non-linear submodel of the form $\beta = \psi(\alpha)$. Then the ML-estimate is approximately

$$(\hat{\alpha}, \hat{\theta}) \approx \arg\min_{(\alpha,\theta)} q(\psi(\alpha), \theta)$$

This non-linear optimization is easier carried out in practice through the $\beta$-profile of (37):

$$\hat{\alpha} = \arg\min_{\alpha} q_{prof}(\psi(\alpha))$$

where

$$q_{prof}(\beta) = \inf_{\theta} q(\beta, \theta) = (\beta - \hat{\beta})^t (H_{\beta\beta} - (H_{\beta\beta} G (H_{prof} + G^t H_{\beta\beta} G) G^t H_{\beta\beta}))(\beta - \hat{\beta})$$

which is obtained from the formula of the marginal precision.
We will consider an application of this technique in *Paper III* (page 57).

### 7.5 Goodness of fit

Consider a realization from the LGCP given by an observation $x$ and hidden log-intensity $\eta$. If we knew the un-observed random variables $\eta$ we would be able to validate the model in two steps: (1) Check that the distribution of $\eta$ is $N(\mu, \Sigma)$. (2) Check that the conditional distribution of $x$ given $\eta$ is $Pois(e^{\eta})$.
As we do not observe $\eta$ in practice it is natural to base goodness of fit assessment on the predictions $\hat{\eta}(x)$.
The Gaussian posterior approximation is

$$\eta|x \sim N(\hat{\eta}_{\theta,\beta}(x), \left(Q_\theta + diag(\hat{\lambda}_\theta(x))\right)^{-1}) \tag{38}$$

where $\hat{\eta}_{\theta,\beta}(x) = \arg\min_\eta l(\theta, \beta|\eta, x)$ and $\hat{\lambda}_\theta(x) = \exp(\hat{\eta}(x))$. If approximation (38) is true then the variance of $\hat{\eta}(x)$ must be

$$V[\hat{\eta}(x)] = Q_\theta^{-1} - E\left[(Q_\theta + diag(\hat{\lambda}_\theta(x)))^{-1}\right]$$

By "removing the expectation" we thus obtain an unbiased estimate of $V[\hat{\eta}(x)]$ by $Q_\theta^{-1} - (Q_\theta + diag(\hat{\lambda}_\theta(x)))^{-1}$ and an approximate standardized residual can be constructed by

$$r_1 = \left(Q_\theta^{-1} - (Q_\theta + diag(\hat{\lambda}_\theta(x)))^{-1}\right)^{-\frac{1}{2}} (\hat{\eta}(x) - \mu) \tag{39}$$

We can avoid removing the expectation by drawing an auxiliary variable $u|x \sim N(0, (Q_\theta + diag(\hat{\lambda}_\theta(x)))^{-1})$. Then the (un-conditional) variance of $\hat{\eta}(x) + u$ is

$$V[\hat{\eta}(x) + u] = Q_\theta^{-1}$$

suggesting the standardized residual

$$r_2 = Q_\theta^{\frac{1}{2}}(\hat{\eta}(x) + u - \mu) \tag{40}$$

Personal simulation studies have shown that $r_2^t r_2$ is closer to the theoretical $\chi^2$-distribution than $r_1^t r_1$. Note that $\hat{\eta}(x) + u - \mu$ is actually an approximate sample from the distribution of $\eta|x$ and therefore assessing the goodness of fit based on $r_2$ follows the line of Waagepetersen (2006). *Paper II* (page 48) provides a simulation experiment of the distribution of $r_2^t r_2$ on a test case of dimension 6000. The $\chi^2$ approximation appears to suffice for this example. If the $\chi^2$ approximation fails an obvious possibility is to simulate the distribution of $r_1$ or $r_2$ directly. This is actually possible even in high dimension because $r_2$ can be calculated using only sparse matrix operations if $Q$ is sparse (*Paper II* (page 48)).

## 7.6 Power

Does the previously introduced standardized residuals have sufficient power to be used for goodness of fit assessment for the LGCP? In this section we consider a small simulation experiment of the spatial LGCP with an exponential correlation structure. The particular case study is based on the model applied in *Paper IV* (page 67) introduced later in this thesis. Five different goodness of fit tests are compared through power simulations.
The test case is specified on a regular $n \times n$-lattice $I_n = \{1, ..., n\}^2$ equipped with the euclidean distance. The covariance of the hidden random field $\eta$ is chosen as $\Sigma = (ae^{-b|i-j|})_{i \in I_n, j \in I_n}$ and a constant mean-log-intensity $\mu$ is imposed at each location. The observation vector $x$ thus have mean $Ex_i = e^{\mu + \frac{1}{2}a}$ and we refer to $\log(Ex_i) = \mu + \frac{1}{2}a$ as the *intercept*. The

parameters of the model are $\theta = (\mu + \frac{1}{2}a, \log a, \log b)$. We choose "true" parameters as

$$\theta_0 = (3, 1, -1)$$

on a regular $20 \times 20$-lattice. These values are inspired by a typical North-Sea case consisting of samples from $\approx 400$ locations with an estimated characteristic distance $b^{-1}$ of $\approx 10 - 20\%$ of the diameter of the area (see *Paper IV*).
For each type of residual-vector $r_1$ (39) and $r_2$ (40) we consider a $\chi^2$-statistic $r^t r$ as well as a Kolmogorov-Smirnov statistic $KS(r)$ given by the uniform distance between the empirical distribution function of $r$ and the standard normal. We also consider using the Laplace approximation of the LGCP-density (22) to measure goodness of fit. The goodness of fit tests are summarized by

1. Chi-square of residuals $r_1^t r_1$.
2. Chi-square of simulation based residuals $r_2^t r_2$.
3. Kolmogorov-Smirnov of residuals $KS(r_1)$.
4. Kolmogorov-Smirnov of simulation based residuals $KS(r_2)$.
5. Laplace approximation of negative log-likelihood $l_x^{LGCP}(\theta_0)$.

Lets now describe how to calculate power functions of the goodness of fit statistics. Note first that each statistic $S$ is a function of the data $x$ and the parameter $\theta_0$. Moreover the simulation based statistics depends on random draws of auxiliary variables $u$. This means that the critical region of a given statistic in the most general setting has the form

$$K(\theta_0) = \{(x, u) : S(x, u, \theta_0) > c\}$$

where $c$ is the $1-\alpha$-quantile of the $P_{\theta_0}$-distribution of $S(x, u, \theta_0)$ in the case of a one-sided test on level $\alpha$. Recall that the power function is given by $\gamma(\theta) = P_\theta(K(\theta_0))$. Practical computation of $\gamma(\theta)$ proceeds as follows:

1. Compute $c$ as the empirical quantile of $S(x, u, \theta_0)$ by simulating 100 draws of data and auxiliary variables $(x, u)$ from the null-model $P_{\theta_0}^{LGCP}$.
2. For each alternative $\theta$ simulate 100 draws of $(x, u)$ from $P_\theta^{LGCP}$ and calculate $\gamma(\theta)$ as the empirical probability that $S(x, u, \theta_0) > c$.

In this study we only consider alternatives within the model structure though the last step could of course be performed for any alternative. For each of the three parameters alternative models are considered by varying the given parameter around its true value keeping the remaining parameters fixed at their true values. Both one-sided and two-sided tests are considered making

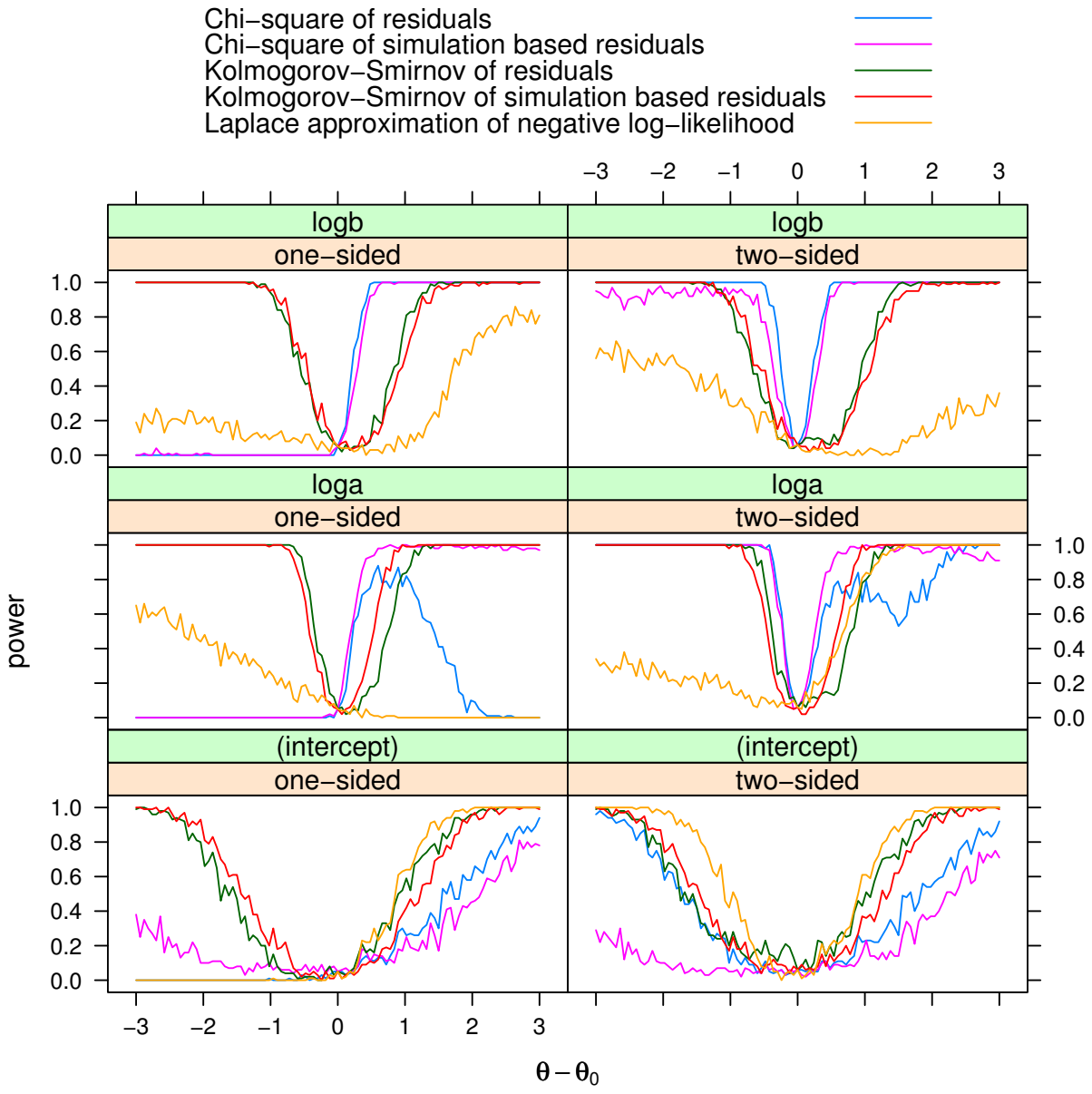



**Figure 5:** Simulation of power function $\gamma(\theta)$ for each of five goodness of fit statistics for the spatial LGCP on a $20 \times 20$-lattice with an exponential covariance structure.

a total of 6 panels (Fig. 5).
The first thing to observe is that three out of the five statistics does not work in the one-sided version namely the $\chi^2$-statistics and the likelihood-statistic. Not surprisingly, the $\chi^2$-statistics are unable to reveal under-dispersion with only large values being critical. This suggests using two-sided versions of the three statistics. In their two-sided versions the $\chi^2$-statistics are superior to the other statistics when it comes to revealing deviations of the covariance-parameters $\log a$ and $\log b$. However, these statistics appears to have very little power as function of the intercept parameter - especially the simulation based $\chi^2$-statistic.
The two-sided likelihood-statistic appears to have most power among all five statistics as function of the intercept-parameter but seems rather useless for revealing deviations of the covariance parameters.
Finally the Kolmogorov-Smirnov statistics appears to work very generally both one-sided and two-sided, however the performance is not impressive. For instance a change in $\log b$ of $\pm\frac{1}{2}$ - corresponding to a 65%-change in the correlation range - is revealed by the Kolmogorov-Smirnov statistics with less than 50% probability. For comparison the two-sided $\chi^2$-statistic reveals this change with a probability close to one.

# 8 LGCP applications

## 8.1 Space-time modelling of length-frequency data

Computational complexity really is an issue when it comes to applying the LGCP on length-based trawl-survey data. For instance a typical survey of cod in the North Sea consists of 400 hauls and 60 length-classes of interest making a total of 24000 random effects. Matrices of this dimension cannot be handled in practice without imposing some special structure.
As previously mentioned the assumption of a sparse precision matrix reduces the computational cost of the Laplace approximation a lot. Rue and Held (2005) establishes the link between sparse precision matrices and Gaussian Markov Random Fields (GMRFs). Can we formulate such GMRF-models to capture relevant heterogeneity of length-based trawl survey data and still obtaining sufficiently sparseness to allow practical application of the method? These are the motivating questions of *Paper II* (page 48).
With focus on a large North-Sea Cod survey we start by formulating a correlation structure inspired by the following considerations

1. Some random parts of the North Sea are more populated than others (large scale spatial correlation)

2. The high and low populated areas may change dynamically - even within a survey (large scale time correlation).

3. Fish swim in small batches with a spatial extension possibly smaller than the dimensions of the trawl and batches have a narrow size composition (small scale size correlation).

4. The trawl is size-selective (size-dependent random thinning).

Assuming separability of "size" and "space-time" we propose the correlation structure

$$\rho(\Delta s, \|\Delta x\|, \Delta t) = \rho_{size}(\Delta s)\rho_{spattemp}(\|\Delta x\|, \Delta t) \tag{41}$$

$$\rho_{spattemp}(\|\Delta x\|, \Delta t) = (1-\nu)e^{-b_1\|\Delta x\|}e^{-b_2\Delta t} + \nu 1_{(\|\Delta x\|=0, \Delta t=0)} \tag{42}$$

of the hidden log-intensity $\eta(s, x, t)$. Here $\|\Delta x\|$ and $\Delta t$ denotes the space and time distance between two samples and $\Delta s$ denotes the separation between two size-classes from each of the samples.
We now turn to the goal of achieving a sparse formulation of the precision. Since all size-classes are represented in each of the samples the covariance takes the form of a Kronecker product.

$$\Sigma = \Sigma_{size} \otimes \Sigma_{spattemp}$$

The Kronecker product is inverted by inverting each factors thus the precision matrix becomes

$$Q = Q_{size} \otimes Q_{spattemp}$$

If one (or both) of the factors have a high proportion of zeros then this will also be the case for $Q$. A simple way to achieve sparseness of $Q_{size}$ is to choose $Q_{size}$ as a band-matrix. For our purpose the precision of a stationary AR(2)-process ($x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \varepsilon_t$, $\varepsilon_t \sim N(0, \kappa^{-1})$) appears to be sufficiently flexible.

$$Q_{size} = \kappa \begin{pmatrix} 1 & -\phi_1 & -\phi_2 & & & & \\ -\phi_1 & \phi_1^2+1 & \phi_1\phi_2-\phi_1 & -\phi_2 & & & \\ -\phi_2 & \phi_1\phi_2-\phi_1 & \phi_2^2+\phi_1^2+1 & \phi_1\phi_2-\phi_1 & -\phi_2 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -\phi_2 & \phi_1\phi_2-\phi_1 & \phi_2^2+\phi_1^2+1 & \phi_1\phi_2-\phi_1 & -\phi_2 \\ & & & -\phi_2 & \phi_1\phi_2-\phi_1 & \phi_1^2+1 & -\phi_1 \\ & & & & -\phi_2 & -\phi_1 & 1 \end{pmatrix} \tag{43}$$

where $\kappa = -\frac{\phi_2-1}{\phi_2^3-\phi_2^2+(-\phi_1^2-1)\phi_2-\phi_1^2+1}$ This precision is defined for $(\phi_1, \phi_2)$ within the triangular region $\{(\phi_1, \phi_2) : \phi_2 > -1,\ \phi_2 < 1+\phi_1,\ \phi_2 < 1-\phi_1\}$. Further sparseness of $Q$ could be obtained by replacing $Q_{spattemp}$ by a 3-dimensional GMRF. This is however somewhat involved because usual constructions of stationary GMRFs are made on regular domains such as the torus or the lattice (Rue and Held, 2005). The highly irregular locations of our space-time coordinates would have to be embedded on a regular 3D-grid e.g. by assigning each coordinate to the nearest grid point (see Rue and Held (2005) page 200). This introduces a new issue of how fine the regular

grid should be. A too rough grid could potentially introduce bias. On the other hand a very fine grid introduces a large number of auxiliary variables (corresponding to the $\eta$s for which no observation is available) making the sparse formulation less beneficial.

## 8.2 Combining LGCP with population model

Section 5.1 provided a mechanistic model of a size-structured population governed by growth, mortality and recruitment. The model was linked to trawl-survey observations through the negative binomial distribution (*Paper I*). The goal of *Paper III* (page 57) is to replace the negative binomial distribution with the LGCP. How does this change affect the information about the population model?
Considering the same data as *Paper I* we start by considering the LGCP with a mean value structure given by (9) and a covariance structure given by (41). This type of model can be fitted using the the methods of section 7.4. Our estimation approach consists of three steps

1. Find approximate ML-estimates $(\hat{\beta}, \hat{\theta})$ of the likelihood (22) using the method described in section 7.4.

2. Construct a second-order expansion $q(\beta, \theta)$ of $-\log L(\beta, \theta)$ around $(\hat{\beta}, \hat{\theta})$ using (37).

3. Fit the size-structured sub-model (13) using the quadratic approximation by writing the sub-model on the form $\beta = \psi(\alpha)$ and optimizing $q(\psi(\alpha), \theta)$ wrt. $(\alpha, \theta)$ and obtain the estimate $(\hat{\alpha}, \hat{\theta}_0)$.

Step 3 replaces the LGCP-likelihood with a quadratic approximation around $(\hat{\beta}, \hat{\theta})$ and has the computational benefit that further model fitting and testing can be carried out without the high-dimensional integrals appearing in the true likelihood function.
The overall conclusion is that temporal changes of the level of the size-spectrum becomes less significant. In particular the catchability can be tested constant which is a major difference compared to *Paper I*.

## 8.3 Applying LGCP to predict abundance surface

Our final application of the LGCP is to predict the abundundance surface of fish. Most prediction methods are based on underlying - more or less transparent assumptions - about the statistical properties of the data under consideration. A correct specification of the statistical model used for predictions is crucial in order to obtain meaningful predictions with realistic uncertainty-estimates. Therefore it seems mandatory to statistically validate the underlying statistical assumptions before applying a given prediction-method.

This is why the LGCP provides an interesting basis for prediction of the log-abundance surface of fish. As a genuine statistical model it allows for validation. Approximate ML-inference can be carried out using the previously described Laplace method. If a given datset passes the goodness of fit validation it makes sense to further apply the statistical model for prediction/interpolation.
*Paper IV* (page 67) is an application of this general approach. Trawl survey data of North Sea cod 1983-2004 are considered. The data constitutes 21 surveys with three age-groups under consideration. A stationary spatial Gaussian random field $\eta$ is applied to describe the hidden log-abundance surface of fish separately for each age-group and survey. The random field is defined by a constant mean $\mu$ and a covariance model given as an exponential structure with a nugget effect:

$$\gamma(\Delta x) = a \exp(-b\|\Delta x\|) + d1_{(\Delta x=0)}$$

This model thus have four parameters $\theta = (a, b, d, \mu)$.
The first important conclusion of *Paper IV* is that the LGCP with the proposed covariance structure adequately describes the spatial heterogeneity of the data as the goodness of fit tests based on the standardized residuals (40) are accepted for all surveys.
Assuming independence between surveys it is further concluded that for any fixed age group the parameters of the covariance $(a, b, d)$ does not change significantly during the entire period. This is remarkable because it means that some basic properties of the local behaviour of the log-abundance surface are invariant.

# Bibliography


Baddeley, A. and Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42. ISSN 1548-7660.

Bagge, O., Thurow, F., Steffensen, E., and Bay, J. (1994). The Baltic cod. *Dana*, 10:1–28.

Bates, D. (2004). Sparse Matrix Representations of Linear Mixed Models. *J. of Computational and Graphical Statistics*.

Bates, D., Maechler, M., and Dai, B. (2008). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-26.

Bertalanffy, L. (1938). A quantitative theory of organic growth (Inquiries on growth laws. II). *Human Biology*, 10(2):181–213.

Bhattacharya, C. G. (1967). A simple method of resolution of a distribution into Gaussian components. *Biometrics*, 23:115–135.

Diggle, P. J. and Ribeiro, P. J. (2006). *Model-based Geostatistics*. Springer. ISBN 0-387-32907-2.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322.

Fournier, D., Hampton, J., and Sibert, J. (1998). MULTIFAN-CL: a length based, age-structured model for fisheries stock assessment, with application to south pacific albacore, thunnus alalunga. *Can. J. Fish. Aquat. Sci.*, 55:2105–2116.

Frøysa, K., Bogstad, B., and Skagen, D. (2002). Fleksibest - an age-length structured fish stock assessment model. *Fish. Res.*, 55:87–101.

Fu, C. and Quinn, T. (2000). Estimability of natural mortality and other population parameters in a length-based model: *Pandalus borealis* in kachemak bay, alaska. *Can. J. Fish. Aquat. Sci.*, 57:2420–2432.

Griewank, A. (2000). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Society for Industrial Mathematics.

Hrafnkelsson, B. and Stefansson, G. (2004). A model for categorical length data from groundfish surveys. *Canadian Journal of Fisheries and Aquatic Sciences*, 61(7):1135–1142.

ICES. (2005). Report of the baltic fisheries assessment working group. CM 2005/ACFM:19.

Macdonald, P. and Pitcher, T. (1979). Age-groups from size-frequency data: a versatile and efficient method of analyzing distribution mixtures. *Journal of the Fisheries Research Board of Canada*, 36:987–1001.

Metz, J. and Diekmann, O. (1986). The dynamics of physiologically structured populations. *Lecture notes in biomathematics*, 68.

Møller, J., Syversveen, A., and Waagepetersen, R. (1998). Log Gaussian Cox processes. *Scand. J. Stat.*, 25:451–482.

Møller, J. and Waagepetersen, R. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Pennington, M. (1996). Estimating the mean and variance from highly skewed marine data. *Fishery Bulletin*, 94(3):498–505.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rue, H. (2005). Marginal variances for Gaussian Markov random fields. *NTNU Statistics Report*.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, H., Martino, S., and Chopin, N. (2007). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Preprint Statistics*.

Rue, H., Steinsland, I., and Erland, S. (2004). Approximating hidden Gaussian Markov random fields. *Journal of the Royal Statistical Society Series B(Statistical Methodology)*, 66(4):877–892.

Schnute, J. and Fournier, D. (1980). A new approach to length frequency analysis: growth structure. *Can. J. Fish. Aquat. Sci.*, 37:1337–1351.

Shepherd, J. G. (1999). Extended survivors analysis: An improved method for the analysis af catch-at-age data and abundance indices. *ICES Journal of Marine Sciences*, 56:584–591.

Skaug, H. and Fournier, D. (2006). Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics and Data Analysis*, 51(2):699–709.

Smith, B., Botsford, L., and Wing, S. (1998). Estimation of growth and mortality parameters from size frequency distributions lacking age patterns: the red sea urchin (*Stronggylocentrotus franciscanus*) as an example. *Can. J. Fish. Aquat. Sci.*, 55:1236–1247.

Sullivan, P. (1992). A Kalman Filter Approach to Catch-at-length Analysis. *Biometrics*, 48:237–257.

von Foerster, H. (1959). *The kinetics of cellular proliferation.*, chapter Some remarks on changing populations., pages 382–407. Grune and Stratton, New York, USA.

Waagepetersen, R. (2006). A Simulation-based Goodness-of-fit Test for Random Effects in Generalized Linear Mixed Models. *Scandinavian Journal of Statistics*, 33(4):721–731.

# Paper I

# How to validate a length-based model of single-species fish stock dynamics

**Kasper Kristensen, Peter Lewy, and Jan E. Beyer**

**Abstract:** This paper validates a new length-based model of the dynamics of fish stocks or crustaceans by hierarchically testing statistical hypotheses and thereby investigating model complexity. The approach is based entirely on scientific survey data and on determination of the statistical distributions of the number of fish caught per haul in each length class. In our example, the negative binomial distribution is statistically accepted and linked to the population level through the new length-based model. The model is derived from the characteristics of continuous recruitment, individually based growth, and continuous, length-dependent mortality rates. Continuous recruitment with annually varying recruitment peaks and individually based growth was crucial for obtaining a model that could be statistically accepted. Natural mortality was estimated as well by the model. The model was applied to survey data for Atlantic cod (*Gadus morhua*) in the Baltic. Its simple generic nature, as well as the validation procedure, is useful in studying and understanding life history and stock dynamics.

**Résumé :** Notre travail valide un nouveau modèle de la dynamique des stocks de poissons basé sur les longueurs en testant des hypothèses statistiques de manière hiérarchique ; il examine ainsi la complexité du modèle. La méthode se base entièrement sur des données d'inventaire scientifique et sur la détermination des distributions statistiques des nombres de poissons capturés par trait de récolte dans chacune des classes de longueur. Nous acceptons dans notre exemple une distribution binômiale n égative et nous la relions au niveau de la population à l'aide du nouveau modèle basé sur les longueurs. Le modèle est dérivé des caractéristiques du recrutement continu, de la croissance individuelle et des taux de mortalité continus en fonction de la longueur. Le recrutement continu avec des pics de recrutement qui varient d'une année à l'autre et la croissance basée sur les individus sont de grande importance pour l'élaboration d'un modèle qui soit statistiquement acceptable. Le modèle estime aussi la mortalité naturelle. Nous avons appliqu é le modèle à des données d'inventaire de la morue franche (*Gadus morhua*) de la Baltique. La nature générique simple et la procédure de validation du modèle le rendent utile pour l'étude et la compréhension des cycles biologiques et de la dynamique des stocks.

[Traduit par la Rédaction]

## Introduction

Validation is a vital part of the modelling process, since it is here the model is confronted with reality. How to make this confrontation in an objective way has often been neglected in studies of fish stock dynamics. In this paper, we present a statistical validation of a new length-based model of the dynamics of fish or crustacean stocks. We restrict attention to a single-species approach that is based entirely on scientific survey data.

The motivation for the study is to increase our ability to make reliable predictions of fish stock dynamics. As a prerequisite, the objective of our modelling is to understand the essence of the information on survey catchability and the vital rates of growth, mortality, and recruitment that is contained in data by validating models of differing complexity. If we can test for model complexity and obtain estimates of vital parameters with confidence limits, then our approach has a promising potential.

Received 19 December 2005. Accepted 27 June 2006. Published on the NRC Research Press Web site at http://cjfas.nrc.ca/ on 27 October 2006.
J19066

**K. Kristensen,[1] P. Lewy, and J.E. Beyer.** Danish Institute for Fisheries Research, Charlottenlund Castle, DK-2920 Charlottenlund, Denmark.

[1]Corresponding author (e-mail: kkr@dfu.min.dk).

The validation of model complexity is done by testing statistical, hierarchical hypotheses. The basic hypothesis to be tested concerns the statistical distribution of the observed catch per unit of effort (CPUE) by length. Can a Poisson model for random encounters adequately describe the observations, or must we use an overdispersed distribution, such as the negative binomial (NB)? Choosing an adequate distribution is crucial because an inadequate distribution may seriously affect test results and parameter estimates. These problems will be pinpointed. Results of testing have not been reported in previous studies, where the stochastic variations in CPUE by length have been described by the normal (Sullivan 1992), the log-normal (Frøysa et al. 2002; Fu and Quinn 2000), or the conditional multinomial (Frøysa et al. 2002; Schnute and Fournier 1980; Smith et al. 1998) distributions. When a statistical distribution has been accepted, the next step in the hierarchical testing of hypotheses is to test if a comprehensive stock dynamics model can be accepted. If the comprehensive model is accepted, more simple submodels can be tested. To our knowledge, nobody has statistically tested model complexity of the stock dynamics model used.

Available models of length-based stock dynamics applying statistical methods are usually age–length structured. Frøysa et al. (2002) and Schnute and Fournier (1980) combined standard, age-structured stock dynamics models with growth models and applied the derived models to length-based catch ob-

servations, while Sullivan (1992) applied a purely length-based state–space model. Sullivan (1992) and Frøysa et al. (2002) approximated the discrete probabilities that fish either do not grow or grow into the neighbor length intervals for each length group. Other authors directly estimated the parameters of the von Bertalanffy growth equation (VBGE) assuming that mean length-at-age follows a specified distribution around a VBGE curve (Schnute and Fournier 1980; Fournier et al. 1998; Fu and Quinn 2000). These age–length structured approaches include both recruitment and temporal dynamics and are discrete in length and time. However, recruitment is assumed to take place instantaneously the same time each year, disregarding the fact that recruitment generally occurs continuously over time with an annually varying peak. Furthermore, the discretization implies some limitation in model assumptions. For instance, Sullivan (1992) and Frøysa et al. (2002) assumed that the probabilities of growing into neighboring length intervals were constant over time. Even if fish growth remains unchanged, this assumption is violated if the length-specific mortality varies over time. Smith et al. (1998) avoided the problems of discretization by using a continuous, statistically based spectrum model with individual variability in growth and length-dependent mortality. Their approach, however, assumed the stock to be in steady state and was therefore unable to consider temporal changes in mortality and recruitment.

There are two reasons why we developed a new approach to modelling the length-based fish stock dynamics. First, to produce a generic model platform we want to avoid possible bias of discretization (Xiao 2005) by formulating a time- and length-continuous model. Second, we want to include continuous recruitment, individual-based growth, and temporal changes of survey catchability because we want to test the importance of these processes. Such requirements are conveniently dealt within a time- and size-continuous approach (i.e., a size-spectrum model).

In the present paper, model complexity is tested by the likelihood ratio test. Regarding the continuous recruitment model, the timing of the recruitment peak and its variation is allowed to change by year and is estimated by the model. Individual growth is modelled using the VBGE assuming that each individual has its own von Bertalanffy asymptotic size ($L_\infty$). Length-structured models for survey catchability and fishing and particularly natural mortality are also included. The estimability of the parameters is examined by Monte Carlo simulation. Only data from scientific surveys are applied, while fishery data are not included in the analysis. One reason for this is that testing whether a specified distribution adequately describes data requires that several independent observations are available following the same distribution in question. Such observations are only provided from surveys conducted within a short time range and selecting individual hauls randomly in an area. The use of survey data is further relevant when the quality of catch data is poor or when such data are not available. Finally, only length-based data are used, which is especially relevant when the age determination is uncertain. The present length-based model is applied to research survey data for Atlantic cod (*Gadus morhua*) in the eastern Baltic for which both problems apply (Reeves 2003). Survey-based but age-structured methods have also been considered by Cook (1997) and Beare et al. (2005).

## Modelling the number density of a population

In the present model, the life history of an individual is determined by growth and mortality. Each individual is assumed to follow its own growth pattern, while the mortality depends on the size of the individual and the time. Recruitment is assumed to take place continuously in time. However, first we derive the number density in case of discrete recruitment. (Refer to the List of symbols for an explanation of all symbols used.)

Let $R_0$ denote a number of individuals recruited at the same time $t_0$ with the same size $L_0$. The fish growth is modelled using the VBGE. All individuals are assumed to have the same growth parameter $k$, while $L_\infty$ varies individually following a common distribution with density $u$ on $(L_0, \infty)$. For an individual fish with a given $L_\infty$, the length at time $t$ is

$$L(t, L_\infty) = L(t, L_\infty|k, t_0, L_0) = L_\infty - (L_\infty - L_0)\exp[-k(t - t_0)] \quad (1)$$

At time $t$ the individuals that have size less than $x$ are exactly the ones with an $L_\infty$ belonging to the set

$$\{L_\infty : L(t, L_\infty) \le x\} = [L_0, G(x)] \quad (2)$$

where

$$G(x) = G(x|k, t_0, L_0) = \frac{x - L_0\exp[-k(t - t_0)]}{1 - \exp[-k(t - t_0)]} \quad (3)$$

Let $z(x, t)$ denote the size- and time-dependent total mortality. Then the number of survivors with size contained in the interval $(L_0, x)$ at time $t$ is given by

$$R_0\int_{L_0}^{G(x)} \exp\left\{-\int_{t_0}^{t} z[L(s, L_\infty), s]\,\mathrm{d}s\right\} u(L_\infty)\,\mathrm{d}L_\infty$$

By differentiation with respect to $x$ of this expression, we obtain the number density $n$ based on a group of individuals recruited at the same time:

$$n(x, t) = R_0\exp\left(-\int_{t_0}^{t} z\{L[s, G(x)], s\}\,\mathrm{d}s\right) \times u[G(x)]G'(x)$$

It should be kept in mind that both $L$ and $G$ depend on $k$, $t_0$, and $L_0$. To find the $n$ based on a continuously recruited population, we simply add the corresponding number densities. Let $r(t)$ denote the recruitment rate. Then $n$ at time $t$ is

$$n(x, t) = \int_{-\infty}^{t} r(t_0)\exp\left(-\int_{t_0}^{t} z\{L_{t_0}[s, G_{t_0}(x)], s\}\,\mathrm{d}s\right) \times u[G_{t_0}(x)]G'_{t_0}(x)\,\mathrm{d}t_0 \quad (4)$$

It may be noted that eq. 4 alternatively could have been derived directly from solutions $n(x, t, L_\infty)$ to the size-structured von Foerster equation (von Foerster 1959) for a given value of $L_\infty$ and then mixing all these solutions with respect to the probability density $u(L_\infty)$. We shall refer to eq. 4 as a size-spectrum model.

## Statistical model

### Formulation

The $n$s from the previous section are now used to formulate a statistical model for catch data from hauls at random positions. The individual hauls are numbered by $i \in I$ and the time of the $i$th haul is denoted $t_i$. The set of sampling times is given by $T = \{t_i : i \in I\}$. It is assumed that for each $t \in T$ the set $\{i \in I : t_i = t\}$ has at least two elements (i.e., for each sampling time, we have at least two hauls).

Within each haul, individual fish are measured with an arbitrary accuracy and associated with a corresponding length group $C_j$. Data can be summarized by a matrix $N_{ij}$ of counts for the $j$th length group at the $i$th haul.

We consider a simple statistical model with a corresponding hierarchy of hypotheses ($H_0 \supset H_1 \supset H_2 \supset H_3$).

$H_0$

In our main statistical model, $N_{ij}$ are assumed to have an independent, NB distribution with identical parameters within each sampling time ($t \in T$) and length group ($j$) (i.e., the random variables $\{N_{ij} : t_i = t\}$ are identically distributed with mean $\mu_{t,j}$ and variance $\sigma^2_{t,j}$ ($\sigma^2_{t,j} > \mu_{t,j}$)). To be able to estimate parameters in this model, we need at least two hauls for each sampling time.

The likelihood function is given by

$$L[(\mu_{t,j}), (\sigma^2_{t,j})] = \prod_i \prod_j \frac{\Gamma(N_{ij} + \nu_{t_i,j})}{\Gamma(\nu_{t_i,j})\Gamma(N_{ij}+1)} \times \pi_{t_i,j}^{\nu_{t_i,j}} (1 - \pi_{t_i,j})^{N_{ij}}$$

in terms of the density function of the NB distribution with probability parameter $\pi_{t_i,j} = \mu_{t_i,j}/\sigma^2_{t_i,j}$ and size parameter $\nu_{t_i,j} = \mu^2_{t_i,j}/\left(\sigma^2_{t_i,j} - \mu_{t_i,j}\right)$. Note that for a given sampling time $t$, the maximum likelihood estimate $\hat{\mu}_{t,j}$ is just the group average, while $\hat{\sigma}^2_{t,j}$ can not be written on closed form.

$H_1$

To reduce the number of parameters in the main model, we state the variance structure hypothesis:

$$V(N_{i,j}) = \sigma^2_{t_i,j} = a_{t_i} \mu^{b_{t_i}}_{t_i,j} + \mu_{t_i,j} \tag{5}$$

This variance structure is mainly proposed from empirical investigations (even though we can give a mechanistic argumentation in the case of $b_{t_i} = 2$, which is out of the scope of this paper). Note that the additional $\mu_{t_i,j}$ term ensures that $\sigma^2_{t_i,j} > \mu_{t_i,j}$, which is required by the NB distribution.

$H_2$

The size-spectrum model with time-dependent catchability is simply a mean value hypothesis in the previous models obtained by assuming the variance structure (eq. 5) and expressing the expected CPUE in terms of catchability and number densities:

$$E(N_{i,j}) = \mu_{t_i,j} = \int_{C_j} q_{\boldsymbol{\theta}}(x, t_i) n_{\boldsymbol{\theta}}(x, t_i)\, \mathrm{d}x \tag{6}$$

The relevance of letting the catchability be time-dependent will be justified.

$H_3$

Finally we may consider the time-independent catchability model by assuming that $q(x, t) = q(x)$ in $H_2$:

$$\mu_{t_i,j} = \int_{C_j} q_{\boldsymbol{\theta}}(x) n_{\boldsymbol{\theta}}(x, t_i)\, \mathrm{d}x$$

Setting up a hierarchy of hypotheses like this has several benefits. First of all it enables successive statistical tests to validate the spectra models. But more importantly it helps us to localize model problems. For example, the model $H_2$ includes three different assumptions regarding the statistical distribution, a variance structure, and a mean value structure. If the $H_2$ hypothesis has problems fitting the data, it will be possible to find out which of the three assumptions are critical by considering successive likelihood ratios.

To validate the main model ($H_0$) we use the randomization technique described in Appendix A because the random variables $N_{ij}$ have a discrete distribution (and thus, transforming with its distribution function, does not produce a uniform distribution). To test the hypotheses $H_1$,$H_2$, and $H_3$, we use likelihood ratio testing for composite hypotheses (Rao 1965).

### Parametrization

The catchability function is chosen as a symmetric sigmoid curve multiplied by a fishing power $p_t$.

$$q_{\boldsymbol{\theta}}(x, t) = \frac{p_t}{1 + \exp[-\alpha_t (x - L^s_{50})]} \tag{7}$$

Both $p_t$ and $\alpha_t$ depend on the time $t$ of the survey. $\alpha_t$ thus describes changes in the selection pattern over time, while $p_t$ models changes in the overall catch efficiency in the survey.

The mortality $z$ is split into two components: the natural mortality and the fishing mortality:

$$z(x, t) = M_0 + f(x, t)$$

where $M_0$ is an assumed unknown constant, and $f$ is assumed to split into the product of a piecewise constant function of time and a sigmoid function of size

$$f(x, t) = \frac{1}{1 + \exp[-\beta(x - L^f_{50})]} \sum_{i=1}^{n} F^{(i)}_{\infty} 1_{(\tau_{i-1} < t < \tau_i)}$$

For the distribution of $L_\infty$ ($u$) we use a normal distribution with mean $\mu_{L_\infty}$ and standard deviation $\sigma_{L_\infty}$. When there are almost no observed fish larger than $\sim 1/2\mu_{L_\infty}$ (actually 50 cm in our case), it is impossible to estimate $\mu_{L_\infty}$ and $k$ simultaneously. Therefore, it is assumed that $\mu_{L_\infty}$ is known and a value from Bagge et al. (1994) is used.

New individuals are recruited to the size of $L_0$ continuously in time. It is assumed that the recruitment period for a year class is normally distributed. Hence the recruitment rate is a linear combination of normal components:

$$r(t) = \sum_{y \in Y} R_y \phi_{\mu^{\mathrm{recr}}_y, \sigma^{\mathrm{recr}}_y}(t) \tag{8}$$

where $\phi_{\mu,\sigma}$ is the normal density with mean $\mu$ and standard deviation $\sigma$. The mean recruitment time for cohort $y$ is parameterized as a year $y$ plus a date $\Delta t_y$ (i.e., $\mu_y = y + \Delta t_y$).

We use $L_0 = 1$ cm and $Y = \{1997, ..., 2003\}$ in the present case.

The parameters of the statistical model $H_2$ are summarized in the vector $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \Big(R_{1997}, R_{1998}, R_{1999}, R^*_{2000} = 1, R_{2001}, R_{2002}, R_{2003}, \Delta t_{1997}, ..., \Delta t_{2003}, \sigma^{\text{recr}}_{1997}, ..., \sigma^{\text{recr}}_{2003}, p_t^{\text{survey}}, L_{50}^{\text{survey}}, \alpha_t^{\text{survey}}, k, \mu^*_{L_\infty} = 135, \sigma_{L_\infty}, L_{50}^{\text{fishery}}, \beta^{\text{fishery}}, F_\infty^{(<2001)}, F_\infty^{(2001-2002)}, F_\infty^{(2002-2003)}, F_\infty^{(>2003)}, M_0, a_t^{\text{variance structure}}, b_t^{\text{variance structure}}\Big) \tag{9}$$

An asterisk (*) indicates that the parameter is fixed.

**Identifiability**

Some care is needed to avoid that the spectra models $H_2$ and $H_3$ get overparameterized. Since the statistical models are determined by the mean values (eq. 6), we need to dispel all obvious parameter bands appearing here. First of all, by inserting eq. 4 in eq. 6, notice the band between the parameter vector $(p_t)$ and the vector of recruitment sizes $(R_y)$; if we multiply and divide the two vectors, respectively, with the same constant, the model is unchanged. This is taken care of by fixing one of the recruitments to 1 (e.g., $R_{2000} = 1$). With this convention, it is possible to estimate the parameters in both statistical models based on the parametrization given in the previous section. This claim has been verified by re-estimating known parameters in simulated data sets.

However these Monte Carlo experiments showed that a large number of the parameters in the model were highly correlated, indicating the need for a reparametrization. Especially high correlations occurred between $a$ and $b$ from the variance structure and among $(p_t)$ from the catchability. The transformation $(\tilde{a}, \tilde{b}) = (\log a + b : \overline{\log \mu}, b)$ made $\tilde{a}$ and $\tilde{b}$ almost uncorrelated. Furthermore, it appeared that the variables $\log p_{t_1}, \log p_{t_2} - \log p_{t_1}, \log p_{t_3} - \log p_{t_2}$, etc. were much less correlated than the vector $(\log p_t)$. Also the choice of reference year class had a major impact on the range of correlations between the parameter estimates.

With the new parametrization, a Monte Carlo re-estimation experiment was carried out. All parameter estimates were plotted in pairs to determine deviations from the asymptotic normal distribution. The plots showed regular ellipses, indicating that the normal approximation applies.

**Model predictions**

Once we have obtained the estimate $\hat{\boldsymbol{\theta}}$, we can compute model predictions of, for example, the relative biomass and length distribution of the commercial catches. Assuming isometric growth (i.e. $W = q_0 L^3$) for some condition factor $q_0$, the total biomass in the system at time $t$ is given by

$$B(t) = q_0 \int_{L_0}^{\infty} x^3 n(x, t)\, \mathrm{d}x \tag{10}$$

Recall that since $R_{2000} = 1$ is fixed, $n(x, t)$ is only known up to a multiplicative constant. Therefore, eq. 10 can only be used to predict the relative biomass in the system, and for that purpose $q_0$ is not needed. We may compute the number density of the fishery catch during the time period $I$ by the formula

$$c_I(x) = \int_I f(x, t) n(x, t)\, \mathrm{d}t \tag{11}$$

This formula makes it possible to compare the catches for a given length group between different time periods.

To compare results with those from age-based models, it is useful to convert length-specific mortalities to age-specific mortalities. For a given cohort, the overall mortality at time $t$ is given by $z_{\text{cohort}}(t) = -N'(t)/N(t)$, where $N(t)$ is the total number of individuals in the cohort at time $t$. Hence

$$z_{\text{cohort}}(t) = \frac{\int_{L_0}^{\infty} z(x, t) n_{\text{cohort}}(x, t)\, \mathrm{d}x}{\int_{L_0}^{\infty} n_{\text{cohort}}(x, t)\, \mathrm{d}x} \tag{12}$$

This equation also applies when replacing $z$ by $m$ or $f$.

## Data

Our data consist of cod catches from 299 selected hauls obtained from the Baltic International Trawl Survey. Only positions inside the International Council for the Exploration of the Sea (ICES) division 25 are considered. All the hauls are taken with TVL-trawl by the Danish vessel *DANA*. The duration of a haul is $\sim$ 30 min.

The survey is performed twice a year — the spring survey, which takes place around 1 March, and the autumn survey, which takes place around 1 November. The actual haul times are distributed over a 1-month interval around these dates, and we associate an average date with each survey — the so-called sampling time.

A brief overview of the data is given in Table 1. The length of each fish has been measured to an accuracy of 1 cm and a length range from 5 to 60 cm is considered. We define the mean CPUE for a given 1 cm length group as the average number of fish caught at a given time for that given length group in a haul. The mean CPUEs per length group are illustrated (Fig. 1), and von Bertalanffy growth curves are also shown (with parameters given in the caption) in order to follow the cohorts through time. The positions of the peaks in the length distributions are reasonably well described by the growth curves. The growth

**Fig. 1.** Catch per unit of effort (CPUE) per 1 cm length group at the nine sampling times and Bertalanffy curves with initial size $L_0 = 1$ cm at 1 August. Each Bertalanffy curve is marked with the two-digit year classes from 1995 to 2003. Except for the initial time $t_0$, all curves have the same parameters $k = 0.12$ year$^{-1}$ and $L_\infty = 135$ cm.

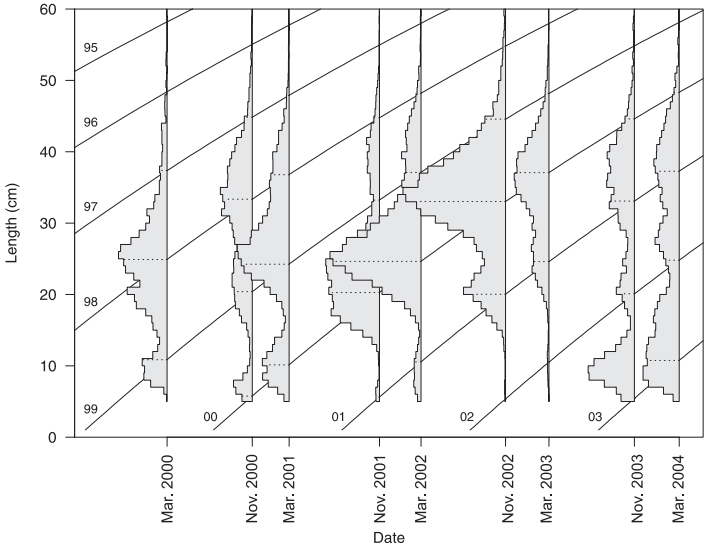



**Table 1.** Number of hauls and fish caught by survey.

| Survey | No. of hauls | No. of fish caught |
|---|---|---|
| Mar. 2000 | 8 | 1 920 |
| Nov. 2000 | 29 | 6 300 |
| Mar. 2001 | 50 | 13 658 |
| Nov. 2001 | 33 | 8 610 |
| Mar. 2002 | 41 | 14 733 |
| Nov. 2002 | 35 | 16 796 |
| Mar. 2003 | 41 | 7 467 |
| Nov. 2003 | 24 | 5 870 |
| Mar. 2004 | 38 | 10 857 |

curves also indicate the extent it is possible to estimate recruitment. It is impossible to estimate the 95 and 96 year classes, as virtually no fish older than 3 years are caught. The year class with the clearest data signal is the 2000 year class; hence it is natural to use the 2000 year class as a reference. All other year classes will be estimated.

## Results

### Validating the main statistical model

The model $H_0$ includes both a distributional assumption and an assumption about independence. These are considered separately. For each of the nine surveys and 55 cm length groups, the maximum likelihood estimates (MLEs) $\hat{\mu}_{t_i}$ and $\hat{\sigma}^2_{t_i}$ have been obtained. By transforming with the estimated distribution function — and randomizing as described in Appendix A — we obtained 299 × 55 residuals ($U_{ij}$), which should follow a uniform distribution on the unit interval. A quantile-quantile (Q-Q) plot indicates that this holds true (Fig. 2*a*). The same residuals ($U_{ij}$) were also plotted against $\mu_{t_i}$ (Fig. 2*b*), showing no systematic patterns.

To demonstrate the importance of choosing a distribution with over-dispersion, the same plots were made with the NB distribution replaced by the Poisson distribution (Figs. 2*c* and 2*d*). The Poisson distribution obviously did not meet the criterion of uniformity.

The model $H_0$ assumes independence between length groups. To validate this assumption, the empirical correlations between length groups were examined, and it appeared that strong correlations existed between neighboring length groups.

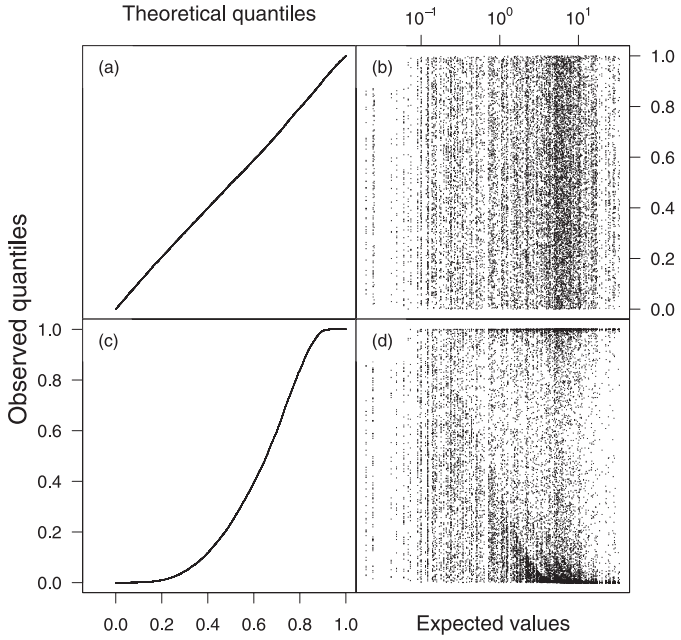
In the light of the length group dependencies, it did not make sense to perform a formal Kolmogorov–Smirnov test on the randomized residuals ($U_{i,j}$) to test whether these residuals followed a uniform distribution. Instead we chose to partition the residuals into corresponding 1 cm length groups (i.e., considering $U_{ij}$ for fixed $j$) and performed the statistical test for each of the length groups in order to test the distributional assumption. Indeed, this lead to acceptance for every group.

We will ignore the length group dependencies throughout this study.

### Examining the variance structure

To reduce the number of parameters in the main model ($H_0$), the variance structure proposed in $H_1$ is suggested, which al-

**Fig. 2.** (*a*) Q-Q plot for the negative binomial (NB) distribution. The observed randomized quantiles are plotted against the theoretical quantiles. (*b*) The observed randomized quantiles for the NB distribution as a function of the estimated mean value parameter. (*c*) Q-Q plot for the randomized quantiles based on the Poisson distribution. (*d*) The observed randomized quantiles for the Poisson distribution as a function of the estimated mean value parameter in the Poisson distribution.




most eliminates half of the parameters. The reasonability of this hypothesis is justified (Fig. 3), showing nine independent analyses — one for each sampling time. The parameter estimates ($\hat{\mu}_{t,j}$ and $\hat{\sigma}^2_{t,j}$) from the full statistical model ($H_0$) are plotted for each sampling time together with the curve $\sigma^2 = \hat{a}_t \mu^{\hat{b}_t} + \mu$, where $\hat{a}_t$ and $\hat{b}_t$ are the MLEs from the variance structure model ($H_1$) at time $t$.

Despite the convincing fits on Fig. 3, the likelihood ratio test of $H_1$ against $H_0$ is rejected. It turns out that the points above the curves generally correspond to fish smaller than 20 cm, while points below the curves correspond to fish larger than 20 cm. This motivates an extension $H_1'$ of $H_1$ with two variance structures for each survey — one for fish smaller than and one for fish larger than 20 cm. It is later shown that $H_1'$ is accepted under $H_0$ (Table 2).

**Spectrum model inference**

Assuming the hypothesis $H_1'$, we now consider the size-spectrum model ($H_2$) with time-dependent catchability. It was assumed that the mean date and the standard deviation of the recruitment process was the same for all year classes earlier than the first sampling time, i.e.:

$$\begin{aligned} &\sigma^{\text{recr}}_{1997} = \sigma^{\text{recr}}_{1998} = \sigma^{\text{recr}}_{1999} \\ &\Delta t_{1997} = \Delta t_{1998} = \Delta t_{1999} \end{aligned} \tag{13}$$

The MLE $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ (eq. 9) was obtained, and the Hessian matrix was checked to be positive definite. Based on $\hat{\boldsymbol{\theta}}$, the expected CPUEs $\mu_{t_i,j}$ were computed and compared with the observed mean CPUEs (Fig. 4). We conclude that the model $H_2$ describes the CPUEs well. Moreover, $H_2$ was accepted by a formal likelihood ratio test (Table 2).

Assuming constant catchability (i.e., time-independent $\alpha_t$ and $p_t$) over time, we estimated parameters based on $H_3$ and obtained a plot equivalent to Fig. 4, which showed that it was impossible for the reduced model to explain the increasing CPUE for the 2000 cohort during year 2002.

To overcome this problem, we also considered a compromise ($H_3'$) between $H_2$ and $H_3$ with time-independent $\alpha_t$ and time-dependent $p_t$. The visual fit was improved compared with $H_3$. However both $H_3$ and $H_3'$ were rejected by the likelihood ratio test (Table 2).

It became apparent that no simple reduction of $H_2$ was possible (Fig. 5). To arrive at a final model, the estimates from $H_2$ were inspected, and the ones that did not differ significantly were collected to form a final hypothesis ($H_2'$):

$$\begin{aligned} &p_4 = p_5 = p_6 \\ &\alpha_1 = \alpha_2 = \alpha_4 = \alpha_5, \quad \alpha_3 = \alpha_6 = \alpha_7, \quad \alpha_8 = \alpha_9 = 0 \\ &\sigma^{\text{recr}}_{1997} = \sigma^{\text{recr}}_{1998} = \sigma^{\text{recr}}_{1999}, \quad \sigma^{\text{recr}}_{2000} = \sigma^{\text{recr}}_{2001} = \sigma^{\text{recr}}_{2002} \\ &\Delta t_{1997} = \Delta t_{1998} = \Delta t_{1999} = \Delta t_{2000} \end{aligned} \tag{14}$$

**Fig. 3.** Maximum likelihood estimates (MLEs; circles) $\hat{\sigma}_i^2$ vs $\hat{\mu}_i$ from the main model ($H_0$). Solid lines indicate the function $\sigma^2 = \hat{a}_t \mu^{\hat{b}_t} + \mu$, where $\hat{a}_t$ and $\hat{b}_t$ are the MLEs from the variance structure model ($H_1$) at time $t$.

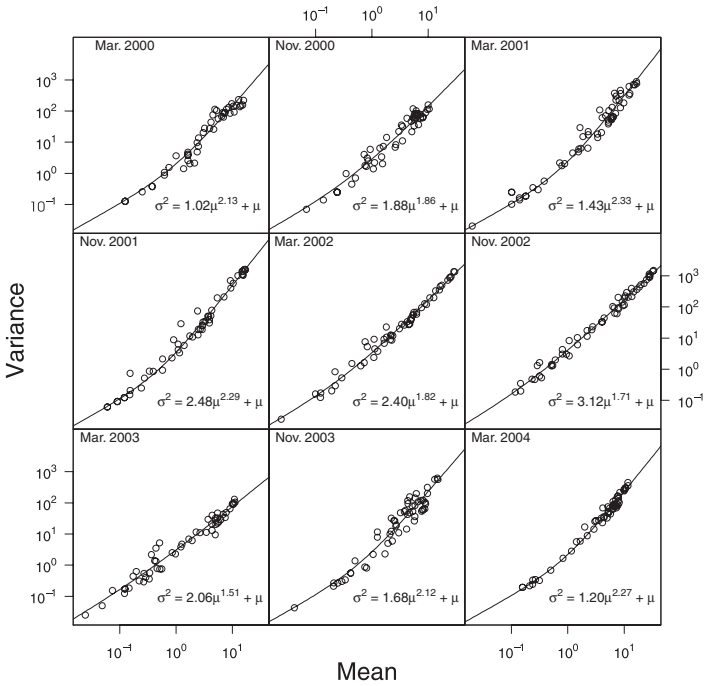



**Table 2.** (*i*) Successive asymptotic likelihood ratio tests. (*ii*) Sequential tests under $H_2$ of $H_3$ (time-independent catchability) and $H_3'$ (time-independent $\alpha_t$ in catchability).

| | | $-\log L$ | $-2\log Q$ | No. of parameters | df | $p$ |
|---|---|---|---|---|---|---|
| (*i*) | $H_0$ | 34 320.0 | — | 990 | — | — |
| | $H_1'$ | 34 565.3 | 490.7 | 531 | 459 | 0.15 |
| | $H_2$ | 34 779.6 | 428.6 | 80 | 451 | 0.77 |
| (*ii*) | $H_3$ | 35 040.5 | 521.7 | 64 | 16 | $< 0.01$ |
| | $H_3'$ | 34 823.9 | 88.6 | 72 | 8 | $< 0.01$ |
| | $H_2'$ | 34 789.2 | 19.3 | 68 | 12 | 0.08 |

The likelihood ratio test of $H_2'$ against $H_2$ was accepted (Table 2). Also the likelihood ratio test of $H_2'$ against $H_0$ supported this conclusion with $p = 0.35$. To validate the applied $\chi^2$ approximation, a simulation study was carried out: 100 data sets were randomly generated from $H_2'$, and the the likelihood ratio test of $H_2'$ under $H_0$ was computed. The simulated distribution agreed perfectly with the theoretical $\chi^2$ distribution with 922 (990 − 68) degrees of freedom.

The final model $H_2'$ had 68 parameters, of which 36 described the variance structure and 32 described the expected CPUEs. A plot of the expected CPUEs did not produce any visible changes from Fig. 4.

The highest absolute parameter correlations in the reduced model were found to be 0.93. However, those correlations involving the mean value parameters were all below 0.90.

The parameters with the highest coefficient of variation (CV) were $\mathrm{CV}(\widehat{M}_0) = 0.66$, $\mathrm{CV}(\widehat{\beta}) = 0.33$, and $\mathrm{CV}(\widehat{\log R_9}) = 0.36$. Almost all other parameters had $\mathrm{CV} < 0.20$.

To test the significance of continuous recruitment versus instantaneous recruitment, the confidence intervals of $\sigma_y^{\mathrm{recr}}$, $y =$ 1997, ..., 2003 were considered. It appeared that none of these included 0. Also, the standard deviation on $L_\infty$ was significantly greater than 0, showing that a model assuming identical growth trajectories for all individuals would be rejected even if individual variability in the spawning time was included.

**Spectra model predictions**

From the parameter estimates based on $H_2$, different kinds of model characteristics were computed — namely the estimated biomass, mortality, and recruitment. Furthermore, the predicted commercial landings were compared with the observed ones.

The biomass relative to year 2002 (i.e., $B(t)/B(2002)$, where $B(t)$ is computed by eq. 10) is shown (Fig. 6*c*). The increasing trend during 2002–2003 was mainly caused by the strong 2000 year class (Fig. 6*b*).

**Fig. 4.** Observed catch per unit of effort (CPUE) and expected CPUE per 1 cm length group at each sampling time obtained from model $H_2$.

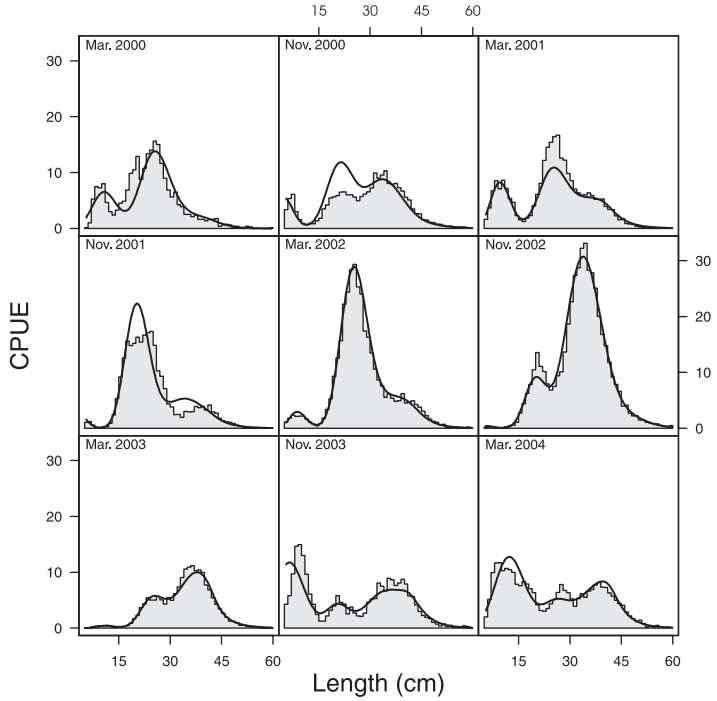
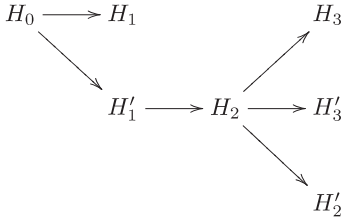



**Fig. 5.** Hypothesis hierarchy. $H_0$, main statistical model without mean and variance assumptions; $H_1$, variance structure model; $H_1'$, extended variance structure model; $H_2$, size-spectrum model with multiplicative time-dependent catchability; $H_3$, time-constant catchability; $H_3'$, time variability in catchability caused by fishing power alone; $H_2'$, final accepted hypothesis obtained by collecting parameters in $H_2$.




No information on the commercial landings has been utilized for model estimation, but catch compositions are available from 2000 until ultimo 2003. The predicted length probability distributions of the yearly commercial landings in that period were computed by eq. 11 and compared with the observed distributions (Fig. 6*d*). The predictions fit the observations quite well for all years. However, when considering the absolute commercial landings over the same period, there is a clear decreasing trend that is not replicated by the model (not shown).

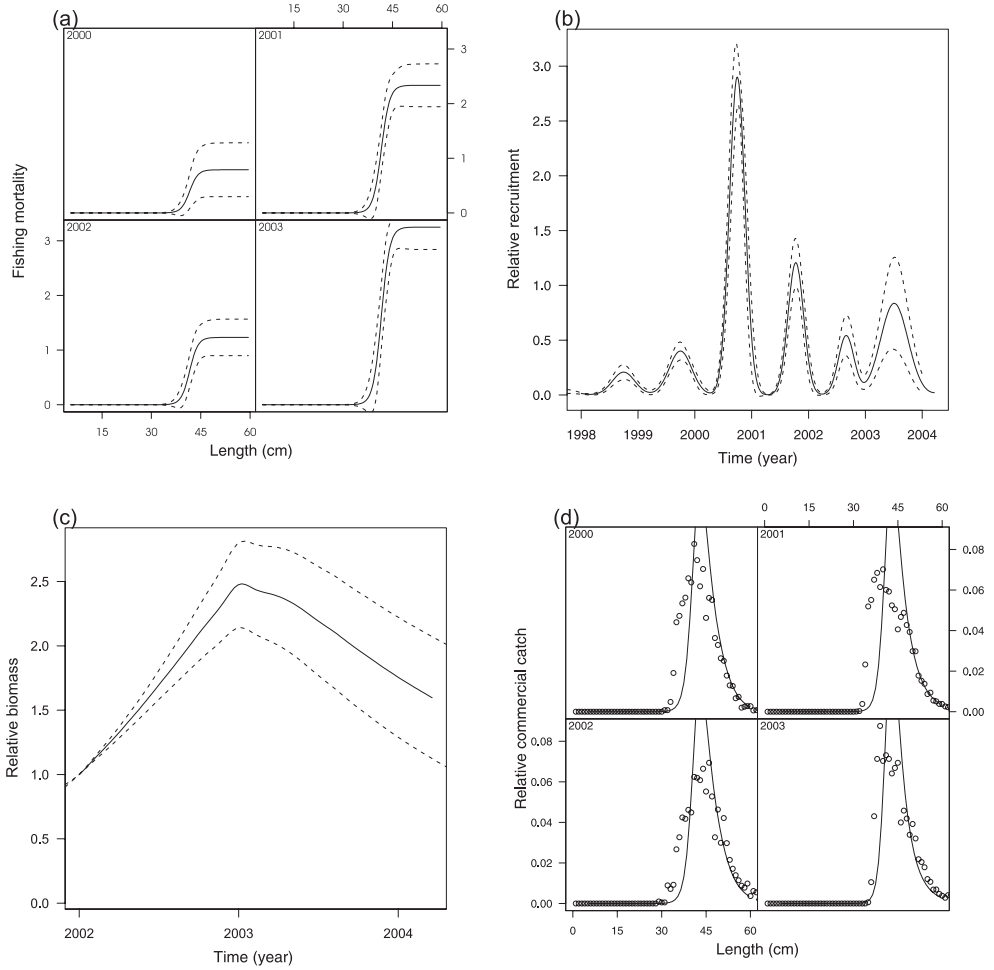
The age-specific fishing mortalities were computed (Table 3) based on eq. 12. These mortalities were slightly lower than the size-specific mortalities (Fig. 6*a*). The reason for this is that only the fraction consisting of the fastest growing individuals of an age group is exposed to the high size-specific mortalities. This phenomenon is maintained by the increasing size dispersion within a cohort implied by the underlying stochastic growth model.

Another consequence of the extreme size-selective mortality is the emergent property that old individuals grow very slowly. This is illustrated by showing the mean of the single cohorts as a function of time (Fig. 7). From this illustration we can conclude that for the typical length distribution of fish within a given survey — which consists of three peaks — the first and second peaks consist primarily of the 0-group and 1-group, respectively, while the third peak consists of all other age groups.

## Discussion

This paper validates a new length-structured model of stock dynamics by testing hierarchically classified hypotheses. Hence,

**Fig. 6.** (*a*) Estimated fishery mortality (solid lines) and 95% confidence limits (broken lines) by fish size. (*b*) Estimated recruitment $R_t$ by year and 95% confidence limits. (*c*) Estimated relative biomass and 95% confidence limits. (*d*) Observed relative length distributions for the commercial catches (circles) and predicted relative length distributions for the fishery based on the spectrum model (solid line).




the approach makes it possible to investigate model complexity, a necessary prerequisite for determining stock dynamics, which is particularly crucial for complex, nonlinear models. Only survey data have been used. Testing hypotheses requires that the applied statistical distribution adequately describes the variation of the observations, and the correspondence between the observations and the distribution should be investigated either graphically or by a formal test. This has been done for observations in the case study here, where the hypothesis of an NB distribution with a specified variance–mean structure was accepted. Testing the significance of a statistical distribution is only possible if several independent, identically distributed observations are available and analysed without prior aggregation as done in the present approach. The NB distribution with an unspecified mean and with the specified variance structure was subsequently used as a general model against which submodels of interest have been tested. The present size-spectrum model is such a submodel specifying the mean of the NB distribution. The spectrum model has been tested against the basic NB model and was accepted. To our knowledge, no previous stock dynamics or assessment approach has formally tested the significance of the model applied.

**Fig. 7.** Estimated mean of the individual cohorts (solid lines) and von Bertalanffy curves from Fig. 1 (broken lines).

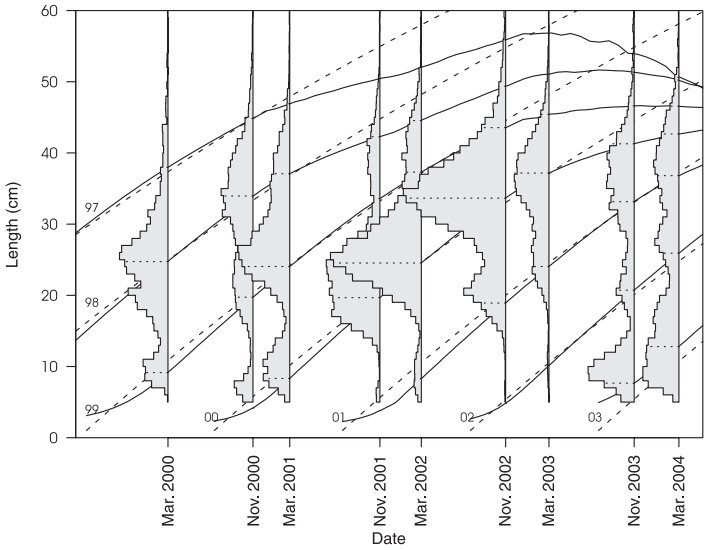



**Table 3.** Age-specific fishing mortality.

| | Age (years) | | | | | | |
|---|---|---|---|---|---|---|---|
| Date | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| Mar. 2000 | 0 | 0 | 0.21 | 0.62 | 0.75 | — | — |
| Nov. 2000 | 0 | 0 | 0.08 | 0.53 | 0.72 | 0.77 | — |
| Mar. 2001 | 0 | 0 | 0.51 | 1.75 | 2.17 | 2.28 | — |
| Nov. 2001 | 0 | 0 | 0.18 | 1.23 | 1.89 | 2.16 | 2.25 |
| Mar. 2002 | 0 | 0 | 0.28 | 0.81 | 1.04 | 1.14 | 1.18 |
| Nov. 2002 | 0 | 0 | 0.09 | 0.75 | 1.04 | 1.11 | 1.15 |
| Mar. 2003 | 0 | 0 | 0.67 | 2.22 | 2.78 | 2.90 | 2.98 |
| Nov. 2003 | 0 | 0 | 0.18 | 1.51 | 2.23 | 2.54 | 2.59 |
| Mar. 2004 | 0 | 0 | 0.60 | 1.79 | 2.19 | 2.34 | 2.33 |

The stock dynamics model developed combines the characteristics of continuous recruitment in time, individual based growth, and continuous, size-dependent mortality rates.As recruitment is assumed to take place continuously, it is possible to test the simpler model of instantaneous recruitment. The model with instantaneous recruitment (with estimation of the optimum time of recruitment) resulted in problems with the interpretation of growth parameters, as the probability of obtaining unrealistic small $L_\infty$ (e.g., $< 20$ cm) was non-negligible. This problem arose because the standard deviation of $L_\infty$ ($\sigma_{L_\infty}$) would compensate for the lack of spawning variation, resulting in too large estimates of $\sigma_{L_\infty}$. These problems were the main reason for the statistical rejection of the spectrum model with instantaneous recruitment. In contrast, the model assuming continuous recruitment did not encounter any of these problems. First, the continuous recruitment model was accepted, and secondly, the distribution of $L_\infty$ can be described by the normal distribution, with negligible possibility of negative values of $L_\infty$ and with a reasonable estimate of $k$. An additional advantage of modelling and estimating recruitment continuously over time by the annual timing of the peak and its temporal variation is the possibility of investigating and studying recruitment processes.

For the model $H_2'$, it was possible to estimate a constant natural mortality, $M_0$. The estimability of $M_0$ is in agreement with Fu and Quinn (2000), suggesting that (in another length-based model for *Pandalus borealis*) constant $M_0$ can be estimated together with a survey catchability varying over time. Assuming, however, that $m(x,t)$ is constant for all fish sizes is of course incorrect, as $m(x,t)$ increases for the smaller fish (ICES 2005*a*).We therefore formulated an alternative, more biologically realistic model by expressing $m(x,t)$ as a size-dependent function, $m(x) = 23\exp(-x+1) + M_\infty$, which is equal to 23 for fish of length 1 cm, decreases to $M_\infty$ when the fish length increases, and is close to $M_\infty$ for $x$ larger than 5 cm. The mortality rate of 23 corresponds to the value used by Bradford (1992) as late cod larvae daily mortality of 0.063. For this model, $M_\infty$ was estimated to 0.16, which is close to and not significantly different from 0.18 previously estimated. The estimates of biomass, recruitment, and fishing mortality are likewise very similar for the two $m(x,t)$ assumptions. Neither did reducing the mortality rate of larval cod by 50% change the results. This suggests that natural mortality can be assumed constant for all sizes when estimating relative biomass, recruitment, and fishing mortality in the size-structured stock dynamics model.

The suggested model considers the individual growth of each fish. The assumption of individual growth patterns was formulated assuming that each fish has its own $L_\infty$ and that the individual values originate from a normal distribution with a mean of 135 cm and a variance to be estimated. The estimated stan-

dard deviation of $L_\infty$ is significantly larger than zero, indicating that the individual growth model is a major improvement compared with a growth model where all fish have the same growth. Individual variability could also be associated with $k$ instead of $L_\infty$, but results of Swain et al. (2003) suggest that for Atlantic cod the model implemented here fits better than the model with varying $k$.

The model is continuous in time and size, which has the advantage that the stock dynamics model is formulated independently of any discretization. Furthermore, observations can be used as a basis for estimation irrespective of the timing, which makes it possible to include additional catch or survey information collected at other times of the year.

The NB distribution was accepted adequately to describe data and was applied further in the calculations. To illustrate consequences of applying a wrong distribution, the Poisson distribution was applied, for which we know that the variance will be underestimated. The MLEs were calculated and the tests of $H_2$ and $H_3$ were carried out. The result was (not surprisingly) that both tests were rejected wrongly, indicating that the spectrum model does not adequately describe stock dynamics. For the $H_2$ model, the MLEs were significantly different from those obtained using the NB distribution.

High correlations were encountered between the number of fish caught per haul in adjacent length groups. Therefore, a multivariate NB distribution should have been used. Unfortunately, such a distribution does not exist. Intuitively, the high correlations should reduce the information in the data, resulting in more uncertain estimates. If the model was able to account for the correlations, we would expect higher test probabilities in general. This could potentially cause simpler model structures to be accepted.

The biomass and mortality estimates of the first 2 years of data are based primarily on assumptions on recruitment, its temporal distribution, and mortality rates for the years prior to the first data year. However, since the stock estimates for the first 2 years are sensitive to the assumptions made, these estimates are more uncertain than those of the following years, an uncertainty that is not reflected in the confidence limits shown. For cases with several years included, this is not so much a problem as it is for the present case, where only few years are included. It is actually surprising that the analysis can be performed based on data for only 5 years.

Estimated fishing mortality for fish larger than 45 cm and fish older than 3 years is very high ($F$ ranges from 1 to 4) compared with the values of $\sim$ 1 estimated by ICES (2005*b*). However, very high mortality rates are supported by the fact that only very few fish larger than 50 cm were caught in the surveys in 2000–2004. To assure that large fish are absent, mortality rates higher than 1 are required. The very high fishing mortality for fish larger than 45 cm, combined with a size selection close to knife-edge selection, results in a mature stock consisting mainly of slow-growing individuals, which have not yet reached a size of about 45 cm. This was indicated by showing that the mean length at age does not increase for fish older than 4 years. It is an open question whether this has led to long-term genetic stock changes.

In conclusion, the present model applied to size-structured scientific survey data is a promising tool to describe and critically examine stock dynamics of a stock for which age determination is uncertain and the quality of catch data is poor. Model complexity can be investigated by testing statistical hypotheses, where different spectra models can be tested thoroughly against a set of probability density distributions describing CPUE by length and haul for each survey. A spectrum model was statistically accepted for which natural mortality and fishing mortality rates, relative biomass and recruitment, and growth were estimated. It is remarkable that the approach reasonably reproduces the relative length distribution of the commercial landings without using these data. The model estimates of fishing mortality could potentially be improved by including commercial catch data by length. This may further enable the estimation of time-varying natural mortality rates.

## Acknowledgements

We thank Anders Nielsen and Holger Hovgaard for useful discussions. The project was funded by the Danish Ministry of Food, Agriculture and Fisheries (the Programme of Development of Sustainable and Selective Fishery). Anna Rindorf provided comments on an earlier version of the manuscript.

## References

Bagge, O., Thurow, F., Steffensen, E., and Bay, J. 1994. The Baltic cod. Dana, **10**: 1–28.

Beare, D., Needle, C., Burns, F., and Reid, D. 2005. Using survey data independently from commercial data in stock assessment: an example using haddock in ICES divison via. ICES J. Mar. Sci. **62**: 996–1005.

Bradford, M. 1992. Precision of recruitment predictions from early life stages of marine fishes. Fish. Bull. (U.S.), **90**: 439–453.

Cook, R. 1997. Stock trends in six North sea stocks as revealed by an analysis of research vessel surveys. ICES J. Mar. Sci. **54**: 924–933.

Fournier, D., Hampton, J., and Sibert, J. 1998. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to south pacific albacore, *Thunnus alalunga*. Can. J. Fish. Aquat. Sci. **55**: 2105–2116.

Frøysa, K., Bogstad, B., and Skagen, D. 2002. Fleksibest — an age–length structured fish stock assessment model. Fish. Res. **55**: 87–101.

Fu, C. and Quinn, T. 2000. Estimability of natural mortality and other population parameters in a length-based model: *Pandalus borealis* in Kachemak Bay, Alaska. Can. J. Fish. Aquat. Sci. **57**: 2420–2432.

ICES 2005*a*. Report of the baltic fisheries assessment working group. ICES CM 2005/ACFM:19.

ICES 2005*b*. Report of the study group on multispecies assesment in the Baltic. ICES CM 2005/H:06.

Rao, C.R. 1965. Composite hypotheses. Chp. 6, sect. e. *In* Linear statistical inference and its applications. John Wiley & Sons, Inc., New York.

Reeves, S.A. 2003. A simulation study of the implications of age-reading errors for stock assessment and management advice. ICES J. Mar. Sci. **60**: 314–328.

Schnute, J., and Fournier, D. 1980. A new approach to length frequency analysis: growth structure. Can. J. Fish. Aquat. Sci. **37**: 1337–1351.

Smith, B., Botsford, L., and Wing, S. 1998. Estimation of growth and mortality parameters from size frequency distributions lacking age patterns: the red sea urchin (*Strongylocentrotus franciscanus*) as an example. Can. J. Fish. Aquat. Sci. **55**: 1236–1247.

Sullivan, P. 1992. A Kalman filter approach to catch-at-length analysis. Biometrics, **48**: 237–257.

Swain, D., Sinclair, A., Castonguay, M., Chouinard, G., Drinkwater, K., Fanning, L., and Clark, D. 2003. Density-versus temperature-dependent growth of Atlantic cod (*Gadus morhua*) in the Gulf of St. Lawrence and in the Scotian Shelf. Fish. Res. **59**: 327–341.

von Foerster, H. 1959. Some remarks on changing populations. *In* The kinetics of cellular proliferation. Grune and Stratton, New York. pp. 382–407.

Xiao, Y. 2005. Catch equations: restoring the missing terms in the nominally generalized Baranov catch equation. Ecol. Model. **181**: 535–556.

## List of symbols

| | |
|---|---|
| $x$ | General notation for size |
| $t$ | General notation for time |
| $k$ | von Bertalanffy growth parameter |
| $L_0$ | Recruitment size |
| $L_\infty$ | von Bertalanffy asymptotic size |
| $L(x, L_\infty)$ | von Bertalanffy growth trajectory |
| $u(L_\infty)$ | Density function describing the probability that an individual is assigned a given $L_\infty$ at the time of recruitment |
| $\mu_{L_\infty}$ | Mean value of $L_\infty$ |
| $\sigma_{L_\infty}$ | Standard deviation of $L_\infty$ |
| $z(x,t)$ | Total mortality as function of size and time |
| $m(x,t)$ | Natural mortality as function of size and time |
| $M_0$ | Parameter in natural mortality |
| $f(x,t)$ | Fishing mortality as function of size and time |
| $F_\infty$ | Parameter vector in fishing mortality |
| $\beta$ | Parameter in fishing mortality |
| $L_{50}^f$ | Parameter in fishing mortality |
| $\boldsymbol{\tau}$ | Vector determining the piecewise constant levels in fishing mortality as a function of time |
| $r(t)$ | Recruitment rate with the property that $r(t)\,\mathrm{d}t$ is the approximate number of individuals recruited to the minimum size $L_0$ during the time interval $(t, t+\mathrm{d}t)$ |
| $n(x,t)$ | Number density function with the property that the number of individuals with size in $(x_1, x_2)$ at time $t$ is given by $\int_{x_1}^{x_2} n(x,t)\,\mathrm{d}x$ |
| $\boldsymbol{\theta}$ | Vector containing all model parameters in a given model. Used as subscript (e.g., $n_{\boldsymbol{\theta}}(x,t)$) to indicate that the function $n(x,t)$ contains unknown parameters |
| $I$ | Set of haul indices |
| $T$ | Set of sampling times |
| $Y$ | Set of recruitment year classes |
| $N_{ij}$ | Matrix of observed number of fish for haul index $i \in I$ and size group $j \in J$ |
| $\mu_{t,j}$ | Expected number of fish in length group $j$ in a haul taken at survey time $t \in T$ |
| $\sigma^2_{t,j}$ | Variance of the number of fish in length group $j$ in a haul taken at survey time $t \in T$ |
| $a_t$ | Variance structure parameters indexed by survey time $t \in T$ |
| $b_t$ | Variance structure parameters indexed by survey time $t \in T$ |
| $R_y$ | Total recruitment for year class $y$ |
| $\mu_y^{\mathrm{recr}}$ | Mean recruitment time for year class $y$ |
| $\Delta t_y$ | Date of the mean recruitment time for year class $y$ (i.e., $\Delta t_y = \mu_y^{\mathrm{recr}} - y$) |
| $\sigma_y^{\mathrm{recr}}$ | Standard deviation of the recruitment rate for year class $y$ |
| $q(x,t)$ | Catchability function |
| $p_t$ | Catchability parameter for $t \in T$ |
| $\alpha_t$ | Catchability parameter for $t \in T$ |

## Appendix A.

**Randomization**

Let $N$ be a discrete random variable on $\mathbb{N}_0$ with distribution function $F$ and let the conditional distribution $U|N = n$ be uniform on the interval $[F(n-1), F(n)]$. Then the distribution of $U$ is uniform on [0,1].

To prove this, let $g$ be the density of $U$. The conditional density of $U|N = n$ is given by

$$g(u|n) = \frac{1}{P(N=n)} 1_{\{u \in [F(n-1), F(n)]\}}$$

Hence the unconditional density is

$$\begin{aligned} g(u) &= \sum_{n=0}^{\infty} g(u|n) P(N=n) \\ &= \sum_{n=0}^{\infty} 1_{\{u \in [F(n-1), F(n)]\}} \\ &= 1 \end{aligned}$$

for any $u \in [0,1]$.

Thus to test whether $N$ has distribution function $F$, we should simulate a random variable $U$ with a uniform distribution on $[F(N-1), F(N)]$ and then test whether $U$ is uniform on [0,1].

**Computational methods**

All computations have been carried out using R (R Development Core Team 2005). An open source R package has been developed for the purpose (available by contacting the authors). The package calls external C++ code, which takes advantage of the free package CppAD (Bell 2005) to evaluate analytical first- and second-order derivatives efficiently.

## Appendix references

Bell, B. 2005. CppAD: a package for C++ algorithmic differentiation [online]. Available from http://www.coin-or.org/CppAD [accessed 1 December 2005; updated 19 October 2006].

R Development Core Team 2005. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available from http://www.R-project.org [accessed 1 December 2005; updated 19 October 2006].

# Paper II

Biometrics , 1–8
In preparation

# Spatio-temporal modelling of population size-composition with the log Gaussian Cox process using trawl survey data

**Kasper Kristensen**
Danish Institute for Aquatic Resources, Charlottenlund Castle, DK-2920 Charlottenlund, Denmark
*email:* kkr@aqua.dtu.dk

Summary: The Log Gaussian Cox Process (LGCP) is a natural and consistent modelling platform to describe spatial point patterns of fish. Most fishery data comes from bottom trawl surveys which may be viewed as random thinnings within rectangular observation windows of a spatial point pattern where only the total number of points within a window is observed.
In this paper the LGCP is considered within the framework of generalized linear geostatistical models (GLGMs). It is described how to perform approximate ML-estimation for models containing a large number of random effects and fixed effects assuming that a sparse formulation of the inverse covariance is feasible.
The approach is applied on a bottom-trawl survey in the North-Sea. A covariance structure is formulated in order to capture the effect of large-scale space-time heterogeneity and small scale size-dependent clustering.
Simulation experiments are conducted to test the method.

Key words: Multivariate Poisson log-normal distribution; Laplace approximation; Log Gaussian Cox Process; Size-correlation; Sparse precision; Spatio temporal modelling.

## 1. Introduction

Modelling of size or age distributions is a key to understanding the population dynamics of fish. Scientific bottom-trawl surveys are conducted to get a snapshot of the size-distribution twice a year in the North-Sea. The statistical interpretation of such data is complicated for several reasons. Typical data are characterized by being over-dispersed multivariate count data with a high proportion of zeros. Heterogeneity on various spatial scales caused by fish schooling and large-scale movement generates patterns in the data. The patterns occur as correlations which - if not accounted for - can lead to over-interpretation and wrong judgement of the uncertainty of the population size-distribution.
The log Gaussian Cox process is a Cox process with random log-intensity following a Gaussian process (Møller et al., 1998). It has successfully been used to model clustering of point patterns caused by environmental heterogeneity. Many ecological models within e.g. forestry or animal breeding are based on point processes (Møller and Waagepetersen, 2004).
A point-process point of view may also be taken for fish population modelling. It is natural to think of a fish population as a heterogeneous spatial point pattern changing dynamically in time. Each point would have an "attribute" in terms of the fish size. Fish samples taken with a trawl would be thought of as a size-dependent random thinning of the point pattern within a rectangular region.
When using the LGCP in practice it is common to discretize the observation window so that the number of points in the discretization cells becomes independent Poisson distributed conditional on a multivariate Gaussian log-intensity (Rue et al., 2007; Brix and Diggle, 2001). This way the LGCP can be put in a context of generalized linear geostatistical models (GLGMs) (Diggle and Ribeiro, 2006).
For trawl-survey data we can treat the haul-rectangles as discretization cells assuming that the log-intensity is approximately constant within a haul-rectangle. This is reasonable from a large scale perspective because the haul-rectangles are small compared to the entire study region.
Various methods have been applied to perform inference for the LGCP comprising Bayesian inference based on MCMC (Møller et al., 1998), Monte carlo maximum likelihood estimation (Møller and Waagepetersen, 2004) and Moment estimation (Brix and Diggle, 2001). Skaug and Fournier (2006) used the Laplace approximation in combination with automatic differentiation to perform approximate ML-estimation. Rue et al. (2007) showed that the Gaussian posterior approximation was sufficiently accurate for inference in many random effects models including the LGCP.
The purpose of the present paper is to describe how to perform approximate ML-estimation for the LGCP within the GLGM setup in cases where the covariance structure has a sparse inverse. The motivation is to be able to handle models with a very large number of random and fixed effects which is necessary in order to apply the model on fishery data.

The approach is illustrated on a single bottom-trawl survey in the North sea by formulating a correlation structure containing the effect of large scale spatio-temporal heterogeneity and small scale size-dependent clustering.
We test the method by simulation and consider goodness of fit assessment.

## 2. Model

### 2.1 *Approximate log Gaussian Cox process likelihood*

In a GLGM context the log Gaussian Cox process is defined as a vector of Poisson counts with a latent log-intensity following a multivariate normal distribution. To write down the joint model of random effects and observations let $\boldsymbol{\eta} \in \mathbb{R}^n$ denote the latent Gaussian random field with mean $\boldsymbol{A\beta}$ being a linear function of a parameter $\boldsymbol{\beta} \in R^k$ where $\boldsymbol{A}$ is the design matrix. The covariance matrix is assumed to be a non-linear function of a parameter vector $\boldsymbol{\theta}$:

$$\boldsymbol{\eta} \sim N(\boldsymbol{A\beta}, \boldsymbol{\Sigma_\theta})$$

The unobserved intensity is

$$\boldsymbol{\lambda} = (e^{\eta_1}, ..., e^{\eta_n})^t$$

Conditional on the intensity the observations are assumed independent Poisson distributed:

$$\boldsymbol{x}|\boldsymbol{\lambda} \sim \otimes_{i=1}^n Pois(\lambda_i)$$

In terms of the precision $\boldsymbol{Q}_\theta = \boldsymbol{\Sigma}_\theta^{-1}$ the full negative log-likelihood is given by

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{x}) = \sum_{i=1}^n e^{\eta_i} - \sum_{i=1}^n x_i\eta_i - \frac{1}{2}\log|\boldsymbol{Q}_\theta| \\ + \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{A\beta})^t\boldsymbol{Q}_\theta(\boldsymbol{\eta} - \boldsymbol{A\beta}) + c \end{aligned} \quad (1)$$

with entire parameter vector $(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $c = \frac{n}{2}\log(2\pi) + \sum_{i=1}^n \log\Gamma(x_i + 1)$. The negative log-likelihood for the observation vector $\boldsymbol{x}$ is obtained by integrating out $\boldsymbol{\eta}$

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{x}) = -\log\int e^{-l(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{x})}\, d\boldsymbol{\eta} \quad (2)$$

The Laplace approximation of this integral is based on a Gaussian posterior approximation (GPA) and has been shown to be sufficiently accurate for many random effects models (Rue et al., 2007; Skaug and Fournier, 2006). For the present case the GPA exists and is unique because the second order derivative of the full negative log-likelihood is everywhere positive definite. It is given by

$$\boldsymbol{\eta}|\boldsymbol{x} \sim N(\hat{\boldsymbol{\eta}}_{\boldsymbol{\beta},\boldsymbol{\theta}}(\boldsymbol{x}), \left(\boldsymbol{Q_\theta} + diag(\hat{\boldsymbol{\lambda}}_{\boldsymbol{\beta},\boldsymbol{\theta}}(\boldsymbol{x}))\right)^{-1}) \quad (3)$$

where $\hat{\boldsymbol{\eta}}_{\boldsymbol{\beta},\boldsymbol{\theta}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{\eta}} l(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{x})$. Before applying the Laplace approximation it is worth applying the GPA on the score function wrt. $\boldsymbol{\beta}$. Replacing $E[\boldsymbol{\eta}|\boldsymbol{x}]$ by $\hat{\boldsymbol{\eta}}_{\boldsymbol{\beta},\boldsymbol{\theta}}(\boldsymbol{x})$ in (A.1) gives

$$\nabla_\beta l(\boldsymbol{\beta}, \theta) \approx -\boldsymbol{A}^t Q_\theta(\hat{\boldsymbol{\eta}}_{\boldsymbol{\beta},\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{A\beta}) \quad (4)$$

Then $\hat{\eta}$ and $\hat{\beta}$ can be found simultaneously by

$$e^{\hat{\boldsymbol{\eta}}} - \boldsymbol{x} + Q_\theta(\hat{\boldsymbol{\eta}} - \boldsymbol{A}\hat{\boldsymbol{\beta}}) = 0 \quad (5)$$

$$\boldsymbol{A}^t Q_\theta(\hat{\boldsymbol{\eta}} - \boldsymbol{A}\hat{\boldsymbol{\beta}}) = 0 \quad (6)$$

using the Newton iterations

$$\begin{aligned} \begin{pmatrix} \boldsymbol{\eta}_{k+1} \\ \boldsymbol{\beta}_{k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\eta}_k \\ \boldsymbol{\beta}_k \end{pmatrix} \\ - \begin{pmatrix} Q_\theta + D_{\boldsymbol{\eta}_k} & -Q_\theta A \\ -A^t Q_\theta & A^t Q_\theta A \end{pmatrix}^{-1} \begin{pmatrix} e^{\boldsymbol{\eta}_k} - x + Q_\theta(\boldsymbol{\eta}_k - \boldsymbol{A\beta}) \\ -\boldsymbol{A}^t Q_\theta(\boldsymbol{\eta}_k - \boldsymbol{A\beta}) \end{pmatrix} \end{aligned} \quad (7)$$

where $D_{\boldsymbol{\eta}} = diag(e^{\boldsymbol{\eta}})$. Each Newton iteration consists of solving a sparse positive definite linear system provided $Q$ and possibly $A$ are sparse. The efficient numerical tool to do this is the sparse Cholesky factorization (Davis, 2006a).

Inserting the corresponding $\hat{\boldsymbol{\beta}}$ in the Laplace approximation of (2) gives an approximate profile likelihood wrt. $\boldsymbol{\theta}$ (up to an additive constant)

$$l_{prof}(\boldsymbol{\theta}|\boldsymbol{x}) \approx l(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\eta}}_{\hat{\boldsymbol{\beta}},\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{x}) + \frac{1}{2}\log\left|\boldsymbol{Q_\theta} + diag(\hat{\boldsymbol{\lambda}}_{\hat{\boldsymbol{\beta}},\boldsymbol{\theta}}(\boldsymbol{x}))\right| \quad (8)$$

This profile can be optimized by using a standard algorithm for non-linear optimization e.g. the BFGS method (Fletcher, 1970). Assuming standard asymptotics for the fixed effects $(\boldsymbol{\beta}, \boldsymbol{\theta})$ the approximate precision of $\hat{\boldsymbol{\theta}}$ is given by the Hessian $\nabla^2_\theta l_{prof}$ of the profile (8). Simultaneous confidence regions of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ can be found by taking derivatives of (4) wrt. $(\boldsymbol{\beta}, \boldsymbol{\theta})$ and using that the hessian of the profile determines the marginal precision of $\theta$:

$$Prec\begin{pmatrix} \hat{\beta} \\ \hat{\theta} \end{pmatrix} = \begin{pmatrix} H_{\boldsymbol{\beta}} & -H_{\boldsymbol{\beta}}G \\ -G^t H_{\boldsymbol{\beta}} & H_{prof} + G^t H_{\boldsymbol{\beta}} G \end{pmatrix} \quad (9)$$

where $H_{\boldsymbol{\beta}} = \nabla^2_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \theta)$ , $G = \nabla_\theta \hat{\boldsymbol{\beta}}$ and $H_{prof} = \nabla^2_\theta l_{prof}$. The gradient $G$ can be found using the implicit function theorem (A.4). The full precision (9) can be used to construct a second-order expansion of the likelihood function (2) around $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{x}) - l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}|\boldsymbol{x}) \approx \frac{1}{2}\begin{pmatrix} \beta - \hat{\beta} \\ \theta - \hat{\theta} \end{pmatrix}^t Prec\begin{pmatrix} \hat{\beta} \\ \hat{\theta} \end{pmatrix}\begin{pmatrix} \beta - \hat{\beta} \\ \theta - \hat{\theta} \end{pmatrix} \quad (10)$$

which may be used to fit and test sub-models about $\boldsymbol{\beta}$ and $\theta$ independent of numerical integration provided that the second-order expansion is sufficiently accurate.

As we need to deal with cases with a high number of hyper parameters we have provided the gradient of (8) in Appendix (4). Apparently the derivative of the log-determinant requires the entire inverse of the precision matrix. However it turns out that the inverse is only needed on the non-zero pattern of the precision. An existing recursive algorithm handles this situation (Rue, 2005; Dahl et al., 2005).

### 2.2 *Goodness of fit*

Our approach to goodness of fit assessment follows the general idea of Waagepetersen (2006). Knowing the pair of random effect and observation $(\boldsymbol{\eta}, \boldsymbol{x})$ it would be easy to validate the model by checking normal and Poisson assumptions separately. By making a single draw $\boldsymbol{\eta}^*$ from the posterior distribution $\boldsymbol{\eta}|\boldsymbol{x}$ then the pair $(\boldsymbol{\eta}^*, \boldsymbol{x})$ has the same distribution as $(\boldsymbol{\eta}, \boldsymbol{x})$. Hence model validation may be based on $(\boldsymbol{\eta}^*, \boldsymbol{x})$. Accurate posterior samples can be very difficult to obtain. In this application we use an approximate posterior sample drawn from the GPA (3). An approximate set of standardized residuals $\boldsymbol{u}$ can be obtained using only sparse matrix operations:

- Draw $\boldsymbol{\eta}^*|\boldsymbol{x}$ from (3).
- Let $LL^t = Q$ and put

$$\boldsymbol{u} = L^t(\boldsymbol{\eta}^* - A\boldsymbol{\beta}) \quad (11)$$

The vector $u$ can be used to visually asses the goodness of fit by plotting against covariates.

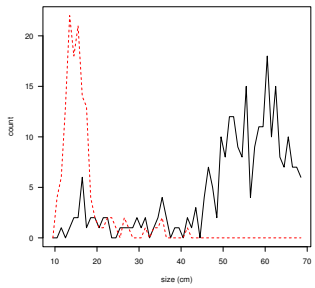



**Figure 1.** Illustration of the small-scale variability in the data: Two samples taken the same day 10 km apart.

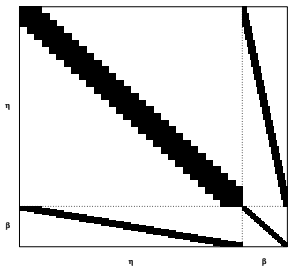



**Figure 2.** Illustration of non-zero pattern of the precision (A.6) (identical to the incidence matrix of the conditional independence graph). A matrix with this pattern must be factorized in each Newton iteration (7).

## 3. Example: Trawl survey data

### 3.1 *Modelling the correlation in time, space, and size*

Bottom trawl surveys in the North sea are conducted twice a year. A typical survey collects fish-samples from approximately 400 different locations covering the entire area. Each fish-sample is a vector of counts representing the number of fish caught in 1 cm size intervals. The samples are taken by 8 different vessels along predetermined routes with approx 10-50 km between successive positions. A survey usually takes 1-2 months.

The purpose of this case study is to apply and validate the LGCP on *one survey*. Our modelling of the hidden log-intensity is based on the following considerations:

(1) Some random parts of the North Sea are more populated than others (large scale spatial correlation)
(2) The high and low populated areas may change dynamically - even within a survey (large scale time correlation).
(3) Fish swim in small batches with a spatial extension possibly smaller than the dimensions of the trawl and batches have a narrow size composition (small scale size correlation).
(4) The trawl is size-selective (size-dependent random thinning).

The hidden log-intensity is modelled by a Gaussian process $\eta(s,x,t)$ indexed by size, space and time where "size" is discrete while "space" and "time" are continuous. It is reasonable to assume that the population size-distribution is unchanged during the relatively short time-period of the survey leading to the assumption that the process should have a size-specific mean $\beta_s = E(\eta(s,x,t))$ which defines the design matrix $\boldsymbol{A}$. Thus $\boldsymbol{\beta}$ has the interpretation of the log-size-distribution of the entire fish-population.

Size-selectivity is easy to model consistently with the LGCP. The correct way to think of size selectivity is as size-dependent random thinning of a point pattern. For a Poisson process this has the effect of downscaling the intensity (prop. 3.7 of Møller and Waagepetersen (2004)) and this can be accomplished by introducing an additive effect to the mean of the log-intensity. In this presentation we do not attempt to explicitly model the size-specific random thinning but just assume that $\boldsymbol{\beta}$ includes this effect.

The residual process $\eta(s,x,t) - \beta_s$ is modelled as a stationary Gaussian process with correlation

$$\rho(\Delta s, \|\Delta x\|, \Delta t) = corr(\eta(s+\Delta s, x+\Delta x, t+\Delta t), \eta(s,x,t))$$

where $\|\Delta x\|$ denotes distance in $km$.

First we attempt to model the spatio temporal Gaussian intensity-landscape for a *fixed size-class*. For simplicity we use a Markovian structure obtained as a product of exponential correlations $e^{-b_1\|\Delta x\|}e^{-b_2\Delta t}$. On top of that process we add Gaussian white noise to model the small scale variability. This has the effect of adding a so called "nugget effect" to the correlation function so that the resulting correlation has the form

$$\rho_{spattemp}(\|\Delta x\|, \Delta t) = (1-p)e^{-b_1\|\Delta x\|}e^{-b_2\Delta t} + p1_{(\|\Delta x\|=0, \Delta t=0)}$$

for $p \in [0,1]$.

According to Fig. 1 extension of this structure to multiple size-classes should require continuity over size of the sample paths which can be achieved by letting

$$\rho(\Delta s, \|\Delta x\|, \Delta t) = \rho_{size}(\Delta s)\rho_{spattemp}(\|\Delta x\|, \Delta t) \qquad (12)$$

Finally we need to choose $\rho_{size}$. As our data has the same size-classes represented for each point in space and time the covariance matrix takes the form of a Kronecker product

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{size} \otimes \boldsymbol{\Sigma}_{spattemp}$$

The Kronecker product is inverted by inverting each factor thus the precision matrix becomes

$$\boldsymbol{Q} = \boldsymbol{Q}_{size} \otimes \boldsymbol{Q}_{spattemp}$$

Choosing $\rho_{size}$ such that $Q_{size}$ is a sparse matrix reduces the computational cost of the Newton iterations (7) dramatically. A sufficiently flexible correlation structure of "size" is obtained by choosing $Q_{size}$ as the precision of a stationary AR(2)-process $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \varepsilon_t$:

$$Q_{size} = \kappa \begin{pmatrix} 1 & -\phi_1 & -\phi_2 & & & & \\ -\phi_1 & \phi_1^2+1 & \phi_1\phi_2-\phi_1 & -\phi_2 & & & \\ -\phi_2 & \phi_1\phi_2-\phi_1 & \phi_2^2+\phi_1^2+1 & \phi_1\phi_2-\phi_1 & -\phi_2 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -\phi_2 & \phi_1\phi_2-\phi_1 & \phi_2^2+\phi_1^2+1 & \phi_1\phi_2-\phi_1 & -\phi_2 \\ & & & -\phi_2 & \phi_1\phi_2-\phi_1 & \phi_1^2+1 & -\phi_1 \\ & & & & -\phi_2 & -\phi_1 & 1 \end{pmatrix} \quad (13)$$

where $\kappa = \frac{1-\phi_2}{\phi_2^3-\phi_2^2-(\phi_1^2+1)\phi_2-\phi_1^2+1}$ and $\varepsilon_t \sim N(0,\kappa^{-1})$.
This precision is defined for $(\phi_1,\phi_2)$ within the triangular region $\{(\phi_1,\phi_2) : \phi_2 > -1,\ \phi_2 < 1+\phi_1,\ \phi_2 < 1-\phi_1\}$.
A preliminary study of the flexibility of the AR(2)-correlation structure showed that only a small part of the triangle gave relevant correlation functions for our applications - more precisely the strip close to the right boundary $\rho_{size}(1) = \frac{\phi_1}{1-\phi_2} \approx 1$. This is due to the obviously high correlation between neighboring size classes (Fig. 1). Reparameterizing to log-distance-to-boundary $\log(1-\phi_1-\phi_2)$ and position-along-boundary $\phi_1-\phi_2$ makes the outer optimization problem much easier for the BFGS-algorithm.
For the present choice of precision matrix $\boldsymbol{Q}$ the non-zero pattern of the system matrix entering in the Newton iterations (7) is shown in Fig. 2. Note that the pattern consists of small dense squares of dimension $400 \times 400$ making the super-nodal variant of the Cholesky factorization suitable (Davis, 2006a).
We end this section by summarizing the parameter vector after convenient changes to the parametrization:

$$\boldsymbol{\theta} = (\log(1-\phi_1-\phi_2), \phi_1-\phi_2, \log b_1, \log b_2, \log\sigma^2, \log(p^{-1}-1)) \quad (14)$$

### 3.2 *Simulation experiment*

Our approach for fitting and validating the LGCP relies on a Gaussian posterior approximation in combination with standard asymptotics. A simulation study is required to test these approximations. The study was based on 100 simulated datasets with "realistic" parameters (actually those obtained by fitting the model to real data (Table 1)). The simulated datasets were based on 200 randomly chosen positions in space and time and 30 size classes.
Coverage of simultaneous confidence regions based on (9) was examined by comparing $\begin{pmatrix}\hat\beta-\beta\\ \hat\theta-\theta\end{pmatrix}^t Prec \begin{pmatrix}\hat\beta\\ \hat\theta\end{pmatrix} \begin{pmatrix}\hat\beta-\beta\\ \hat\theta-\theta\end{pmatrix}$ with the approximate $\chi^2(36)$-distribution (Fig 3a). A similar experiment is shown in (Fig 3b) where $\hat\beta$ is estimated for the *true* $\theta$ now using the conditional precision $H_{\hat\beta}$ from (9) and comparing $(\hat\beta-\beta)H_{\hat\beta}(\hat\beta-\beta)$ with the $\chi^2(30)$-distribution (Fig 3b). Also the coverage of the confidence regions based on the profile likelihood was investigated by comparing $2(l_{prof}(\theta) - l_{prof}(\hat\theta))$ with the theoretical $\chi^2(6)$-distribution. All three comparisons gave a non-significant Kolmogorov-Smirnov p-value. Pairwise plots of parameter estimates looked ellipse-shaped and visualization of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$-parameters showed no sign of bias.
The experiments indicated that (1) Parameters are identifiable (2) The Laplace approximation is sufficiently accurate (3) Standard asymptotics applies with the relatively small amount of data. (4) The parametrization of $\hat\theta$ is suitable.
For each of the simulated datasets we considered the approximate standardized residual-vector $u$ (11) based on the true parameters $(\beta,\theta)$. The sum of squares $u^t u$ were compared with the approximating $\chi^2(6000)$-distribution (Fig. 3d). We also considered the simulated likelihood ratio statistic for the "pure" Laplace method where $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are both treated as fixed effects as opposed to (7). The distribution did not agree well with the $\chi^2(36)$-distribution (not shown). The only difference between the to methods is that the "pure" Laplace method has an extra additive term on the score function (4) which appears when taking the derivative of the Laplace approximation (appendix).

### 3.3 *Application on real data*

We apply the method on the North Sea Cod IBTS survey 1st quarter 2002. The number of samples is 410 and the 1 cm size-classes under consideration are 10-69 cm making a total of 24600 random effects. The dimension of the $\boldsymbol{\beta}$-vector is 60. Estimates of $(\hat\beta,\hat\theta)$ were obtained following the described ML-procedure along with the posterior sample $\eta^*$ and standardized residual vector $u$ (11).
The model was validated by plotting residuals against longitude, latitude, size and time (Fig. 4). No obvious patterns were revealed. A qq-plot of residuals agreed with normal distribution (Fig. 5).
Finally we looked for over-dispersion compared to the Poisson assumption by plotting the observation vector against $e^{\eta^*}$ (not shown). The standard-deviation appeared to approach the mean for large counts consistent with the Poisson-

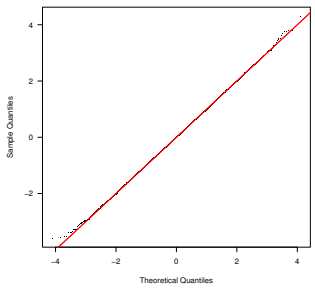



**Figure 5.** qq-plot of standardized residuals obtained by (11)

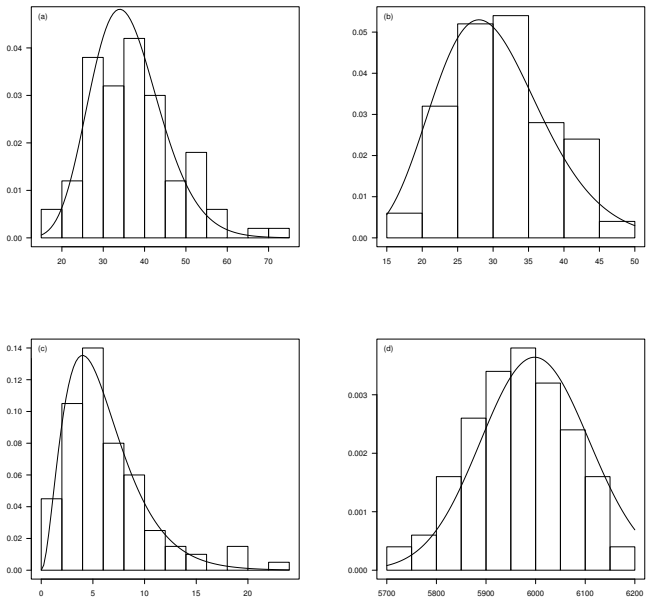



**Figure 3.** 100 simulations. (a) simultaneous deviation of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ from their true values measured in the inner product given by the hessian (9) (histogram) and approximating $\chi^2(36)$-distribution (line). (b) $(\hat{\beta}-\beta)H_{\hat{\boldsymbol{\beta}}}(\hat{\beta}-\beta)$ (histogram) and approximating $\chi^2(30)$-distribution (line). (c) $2(l_{prof}(\theta) - l_{prof}(\hat{\theta}))$ (histogram) and corresponding $\chi^2(6)$-approximation (line). (d) $u^t u$ (11) (histogram) and corresponding $\chi^2(6000)$-approximation (line).

assumption.
Estimated covariance parameters and uncertainties are obtained from the profile likelihood (Table 1).

The corresponding estimates and uncertainties of the size, space and time correlation functions are found using the $\delta$-method (Fig. 6). It appears from Fig. 6c that the time-correlation is approximately constant over a time range of 2 months. This means that the large scale intensity landscape changes rather slowly. A likelihood ratio-test of $b_2 = 0$ gives a p-value of 0.047.
Uncertainties of the log-size composition $\hat{\boldsymbol{\beta}}$ is obtained from the information matrix (9) (Fig. 7). Small scale size-correlation is inherited to the information about population size distribution $\hat{\beta}$ which is not surprising in view of from formula (A.5).
We carried out the exact same analysis for 1st quarter 2001 data. The parameter estimates and uncertainties of correlation parameters were virtually identical to the corresponding results of 2002 indicating robustness of the method. Under a model assuming independence between the two surveys an approximate likelihood ratio test of $\boldsymbol{\theta}_{2001} = \boldsymbol{\theta}_{2002}$ could be constructed based on a quadratic approximation of the likelihood-profile (8) for each survey giving a p-value of 0.40.

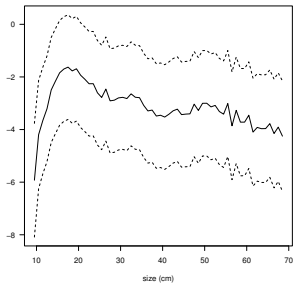



**Figure 7.** Estimated population log-size-composition ($\boldsymbol{\beta}$-parameter) with marginal 95%-confidence intervals.

## 4. Discussion

This paper provides a general procedure for approximate ML-inference for the discretized LGCP put in the context of a generalized linear geostatistical model. The approach is designed for cases involving a large number of observations

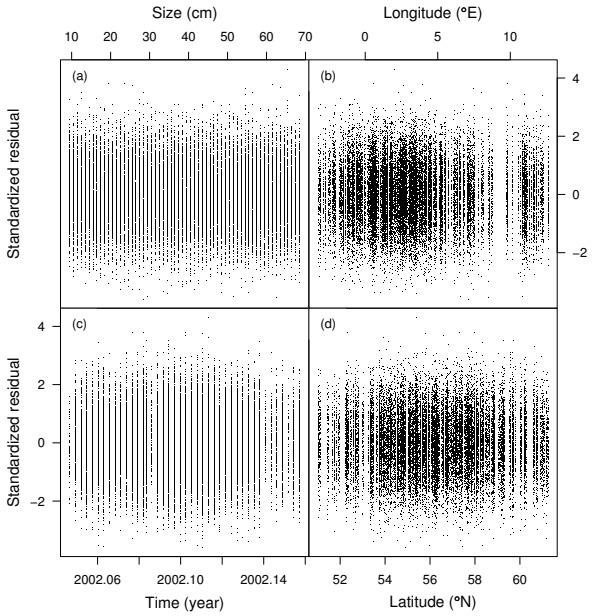



**Figure 4.** Standardized residuals obtained by (11) plotted against (a) size. (b) Longitude. (c) Time. (d) Latitude.

**Table 1**
*Estimated parameters, standard deviations and parameter-correlations.*

| Description | Parameter | Estimate | Sd | Corr | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Size correlation | $\log(1-\phi_1-\phi_2)$ | -4.52 | 0.12 | 1.00 | | | | | |
| Size correlation | $\phi_1-\phi_2$ | 2.56 | 0.06 | -0.64 | 1.00 | | | | |
| Variance | $\log\sigma^2$ | 1.83 | 0.16 | -0.39 | 0.25 | 1.00 | | | |
| Spatial correlation | $\log b_1$ | -5.57 | 0.30 | 0.03 | -0.10 | -0.71 | 1.00 | | |
| Time correlation | $\log b_2$ | 0.01 | 0.80 | -0.05 | -0.03 | -0.26 | 0.33 | 1.00 | |
| Nugget effect | $\log(p^{-1}-1)$ | 1.34 | 0.24 | -0.31 | 0.22 | 0.65 | -0.18 | 0.04 | 1.00 |

and random effects. The joint maximization of random and fixed effects has the interpretation of assigning a flat prior to $\boldsymbol{\beta}$ (appendix) and is thus similar to REML estimation for linear mixed effects models (Jiang, 2006).

To achieve computational speed for large models the approach requires that the covariance has a sparse precision, or equivalently that the latent log-intensity is a Gaussian Markov Random Field (GMRF) (Rue and Held, 2005). Whether this is feasible depends on the application of interest and it is not generally obvious how to achieve this. If the data locations are a subset of a regular grid one can directly apply GMRFs. This was the case for our application on fish-samples from the North Sea. First a covariance structure was formulated to capture relevant heterogeneity caused by large-scale movement and small-scale size-dependent patchiness. Secondly an AR(2)-representation of the size-precision was adopted to obtain the required sparseness.

The space-time precision could have been chosen sparse as well by introducing auxiliary variables (Rue and Held (2005) page 200). It is however not a good idea for the present case where space and time points are highly irregular and relatively few ($\approx 400$). Thus the present formulation does not make any assumptions on the structure of the space time correlation. An exponential covariance with a nugget was chosen for simplicity but more flexible correlations could be tried without much effort like for instance the Matern family (Diggle and Ribeiro, 2006).

Larger and more interesting case-studies would attempt to model many more surveys at once and for such problems a sparse formulation of the space-time precision would be necessary e.g. by using a GMRF on the 3D-torus (Rue and Held (2005) 2.6) and interpolate to irregular grid .

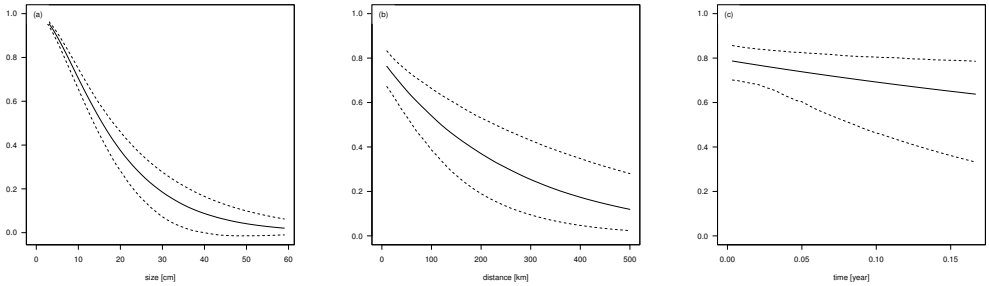



**Figure 6.** Plots of estimated correlation function (12) (line) and 95%-confidence intervals (dashed line) for (a) $\Delta x = 0$, $\Delta t = 0$ as function of $\Delta s$. (b) $\Delta s = 0$, $\Delta t = 0$ as function of $\Delta x$. (c) $\Delta s = 0$,$\Delta x = 0$ as function of $\Delta t$

Unfortunately when combining the 3D-torus with the AR(2) structure of the size precision we get a 4-dimensional graph for which it is less beneficial to apply the sparse Cholesky factorization (Davis, 2006a).

It was shown that the first order Taylor expansion of the scorefunction wrt $(\beta, \theta)$ was quite accurate. For non-linear modelling of $\boldsymbol{\beta}$ it seems obvious to simply replace the likelihood by a quadratic approximation.

Model validation was performed using approximate samples from the posterior distribution $\eta|x$. Attempts were made to improve sampling from the GPA. Metropolis Hastings algorithm was applied with a random walk proposal on the target distribution rescaled to having the identity matrix as second order derivative. After $10^6$ steps the chain had still not converged. The posterior distribution have natural majorizing densities such as the log-gamma distribution and the normal prior distribution. None of these work in practice for rejection sampling in high dimension. It remains an open question how to draw accurate posterior samples for problems of this dimension.

## Acknowledgements

The author thank Professor Håvard Rue for helpful discussions.

## Supplementary Materials

The implementation of the estimation method has been implemented as an R-package (R Development Core Team, 2008) using the sparse matrix library CHOLMOD (Davis, 2006b; Bates and Maechler, 2008). The R-package is available on request.

## References

Bates, D. and Maechler, M. (2008). *Matrix: A Matrix package for R.* R package version 0.999375-14.

Brix, A. and Diggle, P. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63,** 823–841.

Dahl, J., Roychowdhury, V., and Vandenberghe, L. (2005). Maximum likelihood estimation of gaussian graphical models: numerical implementation and topology selection. *UCLA preprint* .

Davis, T. (2006a). *Direct Methods for Sparse Linear Systems*. Society for Industrial Mathematics.

Davis, T. (2006b). User guide for cholmod. Technical report, Tech. rep., University of Florida.

Diggle, P. J. and Ribeiro, P. J. (2006). *Model-based Geostatistics*. Springer. ISBN 0-387-32907-2.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal* **13,** 317–322.

Jiang, J. (2006). *Linear and generalized linear mixed models and their applications*. Springer.

Møller, J., Syversveen, A., and Waagepetersen, R. (1998). Log Gaussian Cox processes. *Scand. J. Stat.* **25,** 451–482.

Møller, J. and Waagepetersen, R. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC.

Neumaier, A. and Groeneveld, E. (1998). Restricted maximum likelihood estimation of covariances in sparse linear models. *Genetics, Selection, Evolution* **30,** 3–26.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rue, H. (2005). Marginal variances for Gaussian Markov random fields. *NTNU Statistics Report* .

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, H., Martino, S., and Chopin, N. (2007). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Preprint Statistics* .

Skaug, H. and Fournier, D. (2006). Automatic approximation

of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics and Data Analysis* **51,** 699–709.

Waagepetersen, R. (2006). A Simulation-based Goodness-of-fit Test for Random Effects in Generalized Linear Mixed Models. *Scandinavian Journal of Statistics* **33,** 721–731.

October 2008

## Appendix

*Computational details*

Here we provide some more details on the calculations of section 2.1. The derivatives of the negative log-likelihood (2) are given by

$$\nabla_{\boldsymbol{\beta}} l(\boldsymbol{\beta},\theta) = -\boldsymbol{A}^t Q_\theta(E_\theta[\eta|x] - \boldsymbol{A\beta}) \tag{A.1}$$

and

$$\begin{aligned}\nabla_{\theta_i} l(\boldsymbol{\beta},\theta) = &-\frac{1}{2}tr(\dot{Q}_\theta Q_\theta^{-1}) + \frac{1}{2}tr(\dot{Q}_\theta V(\eta|x)) \\ &+ \frac{1}{2}(E(\eta|x) - \boldsymbol{A\beta})^t \dot{Q}_\theta (E(\eta|x) - \boldsymbol{A\beta})\end{aligned} \tag{A.2}$$

where $\dot{Q}_\theta$ denotes the elementwise derivative of $\boldsymbol{Q_\theta}$ wrt. $\theta_i$. The joint vector $\hat{\xi}_{\boldsymbol{\theta}} = \begin{pmatrix}\hat{\boldsymbol{\eta}}_{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}\end{pmatrix}$ of random and fixed effects was defined implicitly through the augmented system (5) and (6) of the form

$$f'_{\boldsymbol{\xi}}(\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0$$

where $f(\boldsymbol{\xi},\boldsymbol{\theta})$ is given as the full negative log-likelihood (1). A chain-rule argument on the previous display yields (see also Skaug and Fournier (2006))

$$\nabla_{\boldsymbol{\theta}} \hat{\boldsymbol{\xi}}_\theta = -f''_{\xi\xi}(\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}, \boldsymbol{\theta})^{-1} f''_{\xi\theta}(\hat{\xi}_\theta, \theta) \tag{A.3}$$

which for the present case translates to

$$\nabla_{\theta_i}\begin{pmatrix}\hat{\boldsymbol{\eta}} \\ \hat{\boldsymbol{\beta}}\end{pmatrix} = -\begin{pmatrix}Q_\theta + D_{\hat{\boldsymbol{\eta}}} & -Q_\theta A \\ -A^t Q_\theta & A^t Q_\theta A\end{pmatrix}^{-1}\begin{pmatrix}\dot{Q}_\theta(\hat{\boldsymbol{\eta}} - \boldsymbol{A}\hat{\boldsymbol{\beta}}) \\ -\boldsymbol{A}^t\dot{Q}_\theta(\hat{\boldsymbol{\eta}} - \boldsymbol{A}\hat{\boldsymbol{\beta}})\end{pmatrix} \tag{A.4}$$

used to calculate (9). Taking derivative of (4) gives the remaining part of (9)

$$H_\beta = \nabla^2_\beta l(\beta,\theta) = A^t Q A - A^t Q(Q + D_{\tilde{\boldsymbol{\eta}}})^{-1} Q A \tag{A.5}$$

*Gradient of Laplace approximation*

It is sometimes convenient to give an alternative representation of the model in terms of the augmented vector $\boldsymbol{\xi}$. Consider for some small $\delta > 0$ the augmented positive definite precision matrix

$$\boldsymbol{R_\theta} = \begin{pmatrix}Q_\theta & -Q_\theta A \\ -A^t Q_\theta & A^t Q_\theta A + \delta I\end{pmatrix} \tag{A.6}$$

with determinant $|R_{\boldsymbol{\theta}}| = \delta^k |\boldsymbol{Q_\theta}|$ and consider the LGCP with $\xi \sim N(\boldsymbol{0}, \boldsymbol{R_\theta})$ and no counts associated with the last $k$ entries.

$$l(\boldsymbol{\theta}|\boldsymbol{\xi},\boldsymbol{x}) = \sum_{i=1}^n e^{\xi_i} - \sum_{i=1}^n x_i\xi_i - \frac{1}{2}\log|\boldsymbol{R_\theta}| + \frac{1}{2}\boldsymbol{\xi}^t\boldsymbol{R_\theta}\boldsymbol{\xi} \tag{A.7}$$

Then when $\delta$ approach zero the corresponding full negative log-likelihood (A.7) converge towards (1) except for an additive term $\frac{k}{2}\log\delta$. In conclusion, we can find the gradient of the profile (8) as $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}, \boldsymbol{x})$. Defining

$$h(\boldsymbol{\xi},\boldsymbol{\theta}) = f(\boldsymbol{\xi},\boldsymbol{\theta}) + \frac{1}{2}\log\left|f''_{\boldsymbol{\xi\xi}}(\boldsymbol{\xi},\boldsymbol{\theta})\right|$$

the likelihood profile (8) is $h(\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}, \boldsymbol{\theta})$ and the gradient wrt. $\boldsymbol{\theta}$ is found using the chain-rule (see also Skaug and Fournier (2006))

$$\nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}, \boldsymbol{\theta}) = h'_\theta(\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}, \boldsymbol{\theta}) + h'_{\boldsymbol{\xi}}(\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}, \boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}} \tag{A.8}$$

The derivative $\nabla_{\boldsymbol{\theta}}\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}$ is given by (A.4). The remaining derivatives are now considered. To adapt the notation to the missing Poisson terms of (A.7) corresponding to the last $k$ entries define the intensity and data-vector to be zero for the missing entries i.e. $\lambda_i = e^{\xi_i}$ for $i \leqslant n$ and $\lambda_i = x_i = 0$ when $i > n$. Moreover let $D = diag(\lambda)$. Then the derivatives are given by

$$h'_{\boldsymbol{\xi}}(\boldsymbol{\xi},\boldsymbol{\theta}) = \boldsymbol{\lambda} - \boldsymbol{x} + \boldsymbol{R_\theta}\boldsymbol{\xi} + \frac{1}{2}[\boldsymbol{\lambda}_i((\boldsymbol{R_\theta} + D)^{-1})_{ii}] \tag{A.9}$$

$$h'_{\theta_i}(\boldsymbol{\xi},\boldsymbol{\theta}) = -\frac{1}{2}tr(\boldsymbol{R_\theta}^{-1}\dot{\boldsymbol{R}}_{\boldsymbol{\theta}}) + \frac{1}{2}\boldsymbol{\xi}^t\dot{\boldsymbol{R}}_{\boldsymbol{\theta}}\boldsymbol{\xi} + \frac{1}{2}tr(\dot{\boldsymbol{R}}_{\boldsymbol{\theta}}(\boldsymbol{R_\theta} + D)^{-1}) \tag{A.10}$$

When evaluating expressions like these under the assumption that $\boldsymbol{R_\theta}$ is sparse it is a common trick to note that the inverses $\boldsymbol{R_\theta}^{-1}$ and $(\boldsymbol{R_\theta} + D)^{-1}$ are only required on the non-zero pattern of $\boldsymbol{R_\theta}$ (Rue, 2005; Dahl et al., 2005; Neumaier and Groeneveld, 1998).

# Paper III

# Incorporation of size, space and time correlation into a model of single species fish stock dynamics

**Kasper Kristensen and Peter Lewy**

**Abstract**: This paper improves the statistical interpretation of trawl-survey data by combining the log Gaussian Cox process (LGCP) with a length-based model of single species fish stock dynamics (LBM). The LGCP is suitable for statistical modelling of trawl survey data because of its ability to handle over-dispersed count data with arbitrary correlation structures. A correlation structure is formulated to give a realistic description of the random variation of trawl data caused by spatio-temporal heterogeneity and small scale size-dependent clustering.
We analyze a case study of nine surveys in the Baltic based on the LGCP including size, space and time correlations. The LGCP, fitted following a maximum-likelihood approach, is validated and the biological LBM is statistically accepted as a sub-model. Inclusion of correlations generally results in an information-loss of the size-spectrum level and temporal variations of biological processes becomes less significant. In particular it is shown that by including size, space and time correlations we can statistically accept a hypothesis of constant catchability. The same hypothesis is strongly rejected by a model which ignores the correlations. The interpretation of catchability is crucial for biomass estimates.

## 1. Introduction

Bottom trawl survey data provides statistical information about population dynamics in the sea. Combining a biological length based model with a statistical model of the count data obtained from the survey makes it possible to estimate parameters in the biological model. However the choice of statistical model may greatly affect the outcome of the biological model both in terms of parameter estimates and parameter uncertainties (Deriso et al. 2007) . Also the final choice of complexity of the biological model will depend on the statistical distribution of choice. Therefore it is important to critically validate the statistical distribution before applying it in combination with a biological model.
A recent case study (Kristensen et al. 2006) has shown that by using a negative binomial distribution for the counts obtained from the surveys it is possible to fit a size-based population model surprisingly well based on a fairly small amount of data assuming independence between length-groups. The negative binomial distribution was statistically accepted for each separate length-group. However it was pointed out that the study completely ignored possible dependencies in the data and that this could potentially lead to wrong conclusions.
It is well known that spatial heterogeneity induces dependencies in abundance data. Hence a realistic statistical model should take its starting point by considering the space-time component. In the literature two ways to deal with space-time modelling of fish have been considered: the *mechanistic approach* (Sibert et al. 1999) which attempts to model movement of the individual fish and then scale to population level as opposed to the *geostatistical approach* (Jardim and Ribeiro 2007; Petitgas 2001) which models variations of fish abundance directly on population level by imposing a suitable correlation of the log-abundance surface.
In this paper we follow the geostatistical approach because we are not interested in the spatial component but only in the entire population size-distribution.
A suitable choice of correlation structure makes it possible to account for the fact that the spatial abundance surface of fish changes randomly in both space and time (Petitgas 2001) (spatio-temporal correlation) and at the same time take into account that fish swim in batches of a narrow fish-size composition within batches (small scale size correlation).
Geostatistical modelling of survey data is not straight-forward because survey data are notoriously count data while standard geostatistical models are aimed at normally distributed data. This is why we consider the log Gaussian Cox process (LGCP) (Møller et al. 1998) which is suitable for statistical modelling of over-dispersed count data and at the same time allowing the incorporation of arbitrary correlation structures.
The objective of the present paper is to replace the statistical model applied in Kristensen et al. (2006) with the more realistic LGCP model and to study the biological consequences.
In particular we study catchability assumptions. If a year-class appears to increase from one survey to the next it is common to assume that this is caused by increased catchability. A statistical model which includes space and size-correlations can explain apparent changes in catchability. We illustrate this claim by fitting a size-spectrum model with and without correlations. The likelihood ratio-test of constant catchability is accepted for the model which includes space, time and size-correlations while the same hypothesis is strongly rejected under a statistical model which ignores the correlations.

**K. Kristensen.** Danish Institute for Aquatic Resources, Charlottenlund Castle, DK-2920 Charlottenlund, Denmark
**P. Lewy.** Danish Institute for Aquatic Resources, Charlottenlund Castle, DK-2920 Charlottenlund, Denmark

## 2. Theory

### 2.1. Biological model

The biological model describes the expected size-distribution of a fish population - a so-called size spectrum model. It is obtained by imposing a growth pattern and size dependent mortality to the individuals and then scaling to the population level. Assuming that

1. The growth pattern of an individual follows a von Bertalanffy growth curve (Bertalanffy 1938) with individually varying $L_\infty$ following a probability distribution $u$.
2. Individuals are recruited to the size $L_0$ with recruitment-rate $r(t)$.
3. Each individual is exposed to a size ($s$) and time-specific ($t$) total mortality $z(s,t)$.

it has been shown (Kristensen et al. 2006; Wang and Ellis 2005) that the individual based model scales to the number density

$$n(s,t) = \int_{-\infty}^{t} r(t_0) \exp\left(-\int_{t_0}^{t} z(L_{t_0}(\tau, G_{t_0}(s)), \tau)\, d\tau\right) u(G_{t_0}(s)) G'_{t_0}(s)\, dt_0 \quad [1]$$

where

$$G_{t_0}(s) = \frac{s - L_0 e^{-k(t-t_0)}}{1 - e^{-k(t-t_0)}} \quad [2]$$

and

$$L_{t_0}(t, L_\infty) = L_\infty - (L_\infty - L_0) e^{-k(t-t_0)} \quad [3]$$

For a complete specification of the spectrum model [1] the parametric forms of the functions $r$,$z$ and $u$ must be stated.

**Table 1.** Parameters occurring in [1].

| Symbol | Explanation |
|---|---|
| $k$ | von Bertalanffy growth parameter. |
| $\mu_{L_\infty}$ | Mean value of $L_\infty$. |
| $\sigma_{L_\infty}$ | Standard deviation of $L_\infty$. |
| $M_0$ | Parameter in natural mortality |
| $F_\infty^y$ | Yearly varying asymptotic level of fishing mortality. |
| $\beta^f$ | Parameter in fishing mortality. |
| $L_{50}^f$ | Parameter in fishing mortality. |
| $R_y$ | Total recruitment for year-class $y$. |
| $\mu_y^{recr}$ | Mean recruitment time for year-class $y$. |
| $\sigma_y^{recr}$ | Standard deviation of the recruitment rate for year-class $y$. |

We use the same as Kristensen et al. (2006). The recruitment rate $r$ is chosen as a yearly varying input of Gaussian-shaped peaks . The total mortality $z$ as a sigmoid function of size with a yearly varying asymptote plus an additive level $M_0$ (Appendix [13]) . Finally the distribution of $L_\infty$ was chosen as a normal distribution. All parameters of the size-spectrum model are summarized in Table. 1.

### 2.2. Random intensities

The previously introduced size-spectrum model describes the size-distribution of an entire population. It does not explicitly model where the fish are located. Our approach for distributing the fish in the sea is purely statistical and is based on some rather weak assumptions about the local properties of fish abundance surface: If the abundance is above average at a given location we expect it to be above average at locations nearby. Correspondingly, if a number of fish in one size class is observed to be above average we expect the neighboring size-classes in the same sample to be above average because fish swim in small batches of narrow size-composition.
To meet these requirements let $\eta(s,x,t)$ be a Gaussian stochastic process describing the log-intensity of fish of size $s$ at position $x$ at time $t$. Denote by $\rho(\Delta s, \Delta x, \Delta t)$ the correlation $corr(\eta(s+\Delta s, x+\Delta x, t+\Delta t), \eta(s,x,t))$ assumed only to depend on $(\Delta s, \Delta x, \Delta t)$. The correlation of a Gaussian process is related to the local deviations of the process from its mean because

$$E(\eta(\xi + \Delta\xi | \eta(\xi)) = E(\eta(\xi + \Delta\xi)) + \rho(\Delta\xi)(\eta(\xi) - E(\eta(\xi))) \quad [4]$$

with the notation $\xi = (s,x,t)$.
We apply the structure introduced in Kristensen (2008) given by

$$\rho(\Delta s, \Delta x, \Delta t) = \left((1-\nu) e^{-b_1 \Delta x} e^{-b_2 \Delta t} + \nu 1_{(\Delta x = 0, \Delta t = 0)}\right) \rho_{size}(\Delta s) \quad [5]$$

where $\Delta s$, $\Delta x$ and $\Delta t$ denotes the size, space and time-distance between two samples measured in cm, km and year respectively. The parameter $\nu \in (0,1)$ here denotes the nugget-effect and $\rho_{size}$ is chosen as the correlation of a stationary AR(2)-process with parameters $\phi_1$ and $\phi_2$ (Appendix 6.2).
To understand [5] it is useful to consider the expression when some of the distances are zero. If e.g. $\Delta x = 0$ and $\Delta t = 0$ it means that we are considering a pair of log-intensities corresponding to the same position in space at the same time. The expression reduces to $\rho_{size}(\Delta s)$ which means that $\rho_{size}$ has the interpretation of *small scale size correlation*.
Conversely if $\Delta s = 0$ then it means that we are considering a fixed size-class. The expression reduce to $(1-\nu)e^{-b_1\Delta x}e^{-b_2\Delta t} + \nu 1_{(\Delta x=0, \Delta t=0)}$ which is the correlation of a space time Markov random field modified by adding white noise to model small scale variability.

### 2.3. The LGCP

Length-based survey observations may be organized in a vector of counts $N_i$. With each count $N_i$ is associated a sampling time $t_i$, a position $x_i$, a size $s_i$ and a survey $survey_i$. The LGCP assumes that the counts are independent Poisson distributed conditional on a hidden random log-intensity $\eta_i$:

$$N_i | \eta \sim Pois(\exp(\eta_i))$$

where the hidden log-intensity $\eta$ is assumed to follow a multivariate Gaussian distribution

$$\eta \sim N(\mu, \Sigma)$$

We obtain our statistical model by defining $\Sigma$ in terms of the correlation from the latter paragraph $\Sigma_{i,j} = \sigma^2 \rho(s_i - s_j, x_i - x_j, t_i - t_j)$ and by assuming that the log-intensities have a size and survey specific mean

$$\mu_i = \beta_{size_i, survey_i} \quad [6]$$

The statistical model states that the size-composition of the entire fish-population is unchanged during the survey. This is reasonable as long as the survey duration is short compared to the growth and mortality rates.

### 2.4. Spectrum-model hypothesis

The biological size-based population model is treated as a mean-value hypothesis in the LGCP-model just like in Kristensen et al. (2006)

$$E(N_i) = \int_{C_i} p \,\mathrm{sel}(s) n(s)\, ds \qquad [7]$$

where $\mathrm{sel}(s)$ is an s-shaped size-selectivity function (Appendix [12]) taking values in the interval $(0, 1)$ and $p$ denotes the fishing-power.
For the LGCP [7] is equivalent to a sub-model of [6] given by

$$\beta_{size,survey} = \log\left(\int_{C_{size}} p \,\mathrm{sel}(s) n(s, t_{survey})\, ds\right) - \frac{1}{2}\sigma^2 \qquad [8]$$

because $EN_i = \exp(\mu_i + \frac{1}{2}\sigma^2)$ (Aitchison and Ho 1989). It is convenient to apply the reparameterization $\log\tilde{p} = \log p - \frac{1}{2}\sigma^2$. Then there is no overlap between parameters describing $\mu$ and those describing $\Sigma$.
Kristensen et al. (2006) found it necessary to model $p$ as being survey dependent $p_{survey}$ in order to explain apparently increasing cohorts between surveys. A likely explanation of this phenomenon is that the survey positions (Fig. 1) are concentrated on areas with abundance above average some years and below average other years. It is therefore relevant to test the *constant catchability hypothesis* under a statistical model which accounts for size, space and time correlations in the observations.
All parameters in the size-spectrum model are summarized in the vector $\alpha$ given by

$$\begin{aligned}\alpha = (\quad & R_{1997}, R_{1998}, R_{1999}, R^*_{2000} = 1, R_{2001}, R_{2002}, R_{2003},\\ & \mu^{recr}_{1997}, ..., \mu^{recr}_{2003},\\ & \sigma^{recr}_{1997}, ..., \sigma^{recr}_{2003},\\ & p_t^{(survey)}, L_{50}^{(survey)}, \gamma^{(survey)},\\ & k, \mu^*_{L_\infty} = 135, \sigma_{L_\infty},\\ & L_{50}^{(fishery)}, \delta^{(fishery)},\\ & F_\infty^{(<2001)}, F_\infty^{(2001-2002)}, F_\infty^{(2002-2003)}, F_\infty^{(2003<)},\\ & M_0\quad )\end{aligned} \qquad [9]$$

containing both the parameters of Table 1 and the catchability parameters. A fixed value of $\mu_{L_\infty}$ is used and the recruitment for for year 2000 is fixed to obtain estimability (Kristensen et al. 2006).

### 2.5. Approximate likelihood inference

The likelihood function of the LGCP-model is obtained as a product of a multivariate normal pdf and a Poisson pdf summed over all possible combinations of the un-observed random intensity:

$$L(\beta, \theta) \propto \int_{\mathbb{R}^n} |\Sigma_\theta|^{-\frac{1}{2}} e^{-\frac{1}{2}(\eta - A\beta)\Sigma_\theta^{-1}(\eta - A\beta)} \prod_{i=1}^{n} \frac{e^{\eta_i N_i}}{N_i!} e^{-e^{\eta_i}} d\eta \qquad [10]$$

Here $A$ denotes the design-matrix corresponding to [6] written on vector form $\mu = A\beta$. All covariance parameters are contained in the vector $\theta = (\sigma^2, b_1, b_2, \phi_1, \phi_2, \nu)$. Kristensen (2008) provided an efficient method for estimation in this model which utilizes the fact that $\Sigma_\theta^{-1}$ and $A$ are sparse matrices (Davis 2006a,b). The method is based on a Gaussian approximation of the distribution of $\eta|N$ to compute the integral and is thus similar to the method described in Skaug and Fournier (2006). However we cannot directly apply their approach because of the large number of random effects and fixed effects in our application.
Our estimation approach consists of three steps

1. Find approximate ML-estimates $(\hat{\beta}, \hat{\theta})$ of the likelihood [10] applying a Gaussian posterior approximation as described in Kristensen (2008).
2. Construct a second-order expansion $q(\beta, \theta)$ of $-\log L(\beta, \theta)$ around $(\hat{\beta}, \hat{\theta})$.
3. Fit the size-spectrum model using the quadratic approximation by writing [8] on the form $\beta = \psi(\alpha)$ and optimizing $q(\psi(\alpha), \theta)$ wrt. $(\alpha, \theta)$ and obtain the estimate $(\hat{\alpha}, \hat{\theta}_0)$.

The approximations applied in step 1 and 2 were investigated by simulation in Kristensen (2008). It was concluded that the likelihood function based on a Gaussian posterior approximation was sufficiently accurate to consistently estimate $\theta$ and $\beta$ without any visible bias. Furthermore the distribution of $q(\beta, \theta) - q(\hat{\beta}, \hat{\theta})$ was very close to the theoretical $\chi^2$-distribution indicating that the second order approximation of the likelihood was quite accurate. These conclusions justifies the quadratic approximation applied in step 3. Step 3 replaces the likelihood [10] with a quadratic approximation around $(\hat{\beta}, \hat{\theta})$ and has the computational benefit that further model fitting and testing can be carried out without the high-dimensional integrals appearing in the true likelihood function. One should have in mind that the quadratic approximation only holds in a neighborhood around $(\hat{\beta}, \hat{\theta})$ and thus a parameter estimate $\hat{\alpha}$ obtained using the quadratic approximation can be very different from the true maximum-likelihood estimate if the corresponding $\psi(\hat{\alpha})$ lies outside the neighborhood. However in this case the sub-model would be rejected by a likelihood ratio test and we would discard the estimate anyway.
Confidence regions of parameter estimates are obtained from the observed information matrix (the Hessian) assuming standard asymptotics.
Tests of the spectra-model hypothesis against the unparameterized model are based on the approximate likelihood ratio statistic

$$q(\psi(\hat{\alpha}), \hat{\theta}_0) - q(\hat{\beta}, \hat{\theta}) \sim \chi^2(df) \qquad [11]$$

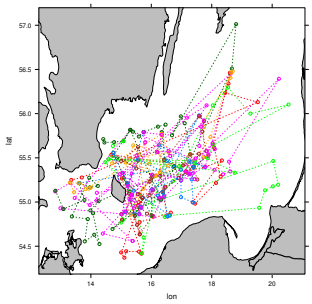



**Fig. 1.** Haul positions of BITS for all 9 surveys (points). The surveys are distinguished by the different colors. A line joins two points if the corresponding hauls are chronological neighbors.

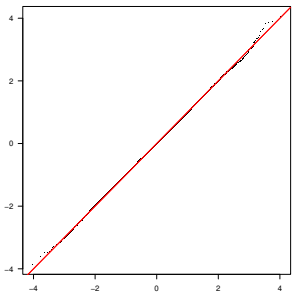



**Fig. 2.** qq-plot of standardized residuals against normal distribution.

where $df$ is the difference between the dimension of $\beta$ and $\alpha$. To validate the model we follow the simulation based approach from Kristensen (2008) : Draw a (high-dimensional) sample $\eta^*$ from the approximate distribution of the random effect $\eta$ given the data $N$. If the model is true then the pair $(\eta^*, N)$ has the same distribution as $(\eta, N)$. So we can asses the goodness of fit of the random effect by checking that $\eta^*$ is normal with mean $\mu$ and covariance $\Sigma$.

## 3. Data

We consider the same data as Kristensen et al. (2006) in order to compare the results of the different approaches. Data consists of 299 hauls from the Baltic International Trawl Survey (BITS) where only positions within ICES subdivision 25 are considered (Fig. 1). Surveys are conducted twice a year during spring and autumn and the duration of a survey is $\approx 1$ month. The present data set includes the 9 surveys from spring 2000 until spring 2004.
The measurement of consideration is the number of fish caught in 59 size-groups in the individual hauls. For each measurement (count) $N_i$ the following co-variates are used: haul identification, longitude, latitude, haul-initialization-time.

## 4. Results

The raw statistical model without any biological assumptions on the mean-value structure [6] was fitted by optimizing the described approximation to [10] with the correlation-structure [5]. Following the simulation based goodness of fit approach we obtained a set of standardized residuals. A qq-plot of the residuals (Fig. 2) agreed with the normal distribution. Plotting the residuals against the covariates size,time, latitude and longitude (Fig. 3) did not reveal any systematic trends.
The ML-estimates of the model are $(\hat{\beta}, \hat{\theta})$ where $\hat{\beta}$ represents the un-parameterized log-size-distribution and $\hat{\theta}$ contains the correlation parameters. Estimates, standard deviations and correlations of the $\theta$-parameters (transformed as in Kristensen (2008) to improve the normal approximation of the likelihood function) are given in table 2 indicating that there was no problem with parameter-identification in this model.

The corresponding correlation functions are shown with (pointwise) confidence limits (Fig. 4). The size-correlation (Fig. 4a) is estimated to be greater than 50% for size-differences less than 20 cm and the effective range - defined as the lag for which the correlation has decreased to 5% - is estimated to be 78 cm (CV=8%). Space and time correlations (Fig. 4b and 4c) are substantial as well with estimated effective ranges of 200 km (CV=16%) and 0.33 year (CV=21%). For comparison the largest spatial distance between two samples is $\sim 500$ km. As the effective range of the time correlation is smaller than 0.5 year there is almost no correlation between two successive surveys. Approximate likelihood ratio tests of (1) ignoring size-correlation ($\phi_1 = \phi_2 = 0$) (2) ignoring space-time correlation ($b_1 = b_2 = \nu = 0$) were both strongly rejected ($p < 10^{-4}$).
A Wald test of using the AR(1)-process to model size-correlation versus the alternative AR(2)-process ($\phi_2 = 0$) was also rejected ($p < 10^{-4}$).
The estimated parameter-vector $\hat{\beta}$ is illustrated with 95%-confidence intervals (Fig. 5). To better understand how the correlation model affects the precision of the $\beta$-parameter we considered a pair of length groups 20 cm and 21 cm for the spring 2001 survey (the survey with the largest number of hauls). 95%-confidence ellipses of the parameter pair $(\beta_1, \beta_2)$ were constructed (Fig. 6) on basis of the observed information matrix for the different combinations of correlation-model. The model assuming independence between all observations has the red circle as confidence band of $(\beta_1, \beta_2)$. Adding space-time correlation (but no size-correlation) results in wider confidence bands (blue circle). However when either of these models are extended to include size-correlation the confidence regions are squeezed to ellipses. It appears from Fig. 6 that the inclusion of size-correlation increases the uncertainty in the direction $y = x$ while the uncertainty in the orthogonal direction is reduced. This observation also holds for other size-classes and generally means that by including size-correlation we increase the information about the log-size-spectrum slope while the information about the overall level of the spectrum is reduced.

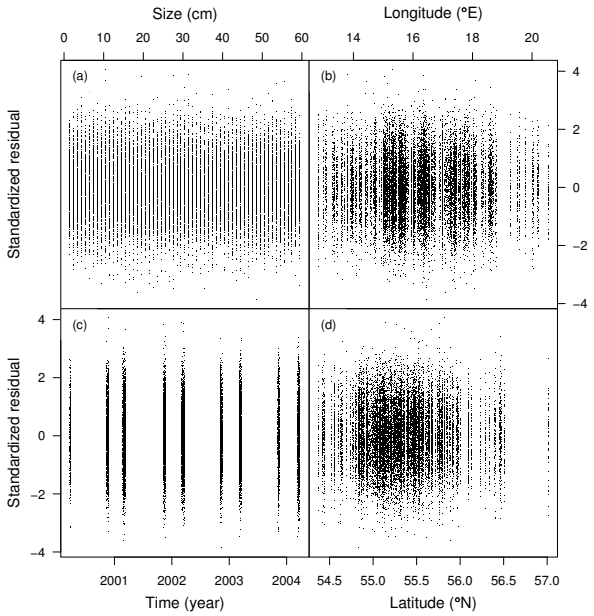



**Fig. 3.** Standardized residuals plotted against (a) size. (b) Longitude. (c) Time. (d) Latitude.

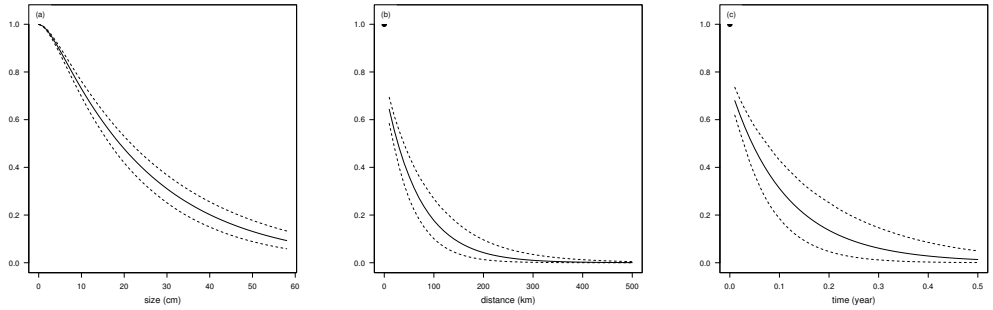



**Fig. 4.** Plots of estimated correlation function [5] (line) and 95%-confidence intervals (dashed line) for (a) $\Delta x = 0$, $\Delta t = 0$ as function of $\Delta s$. (b) $\Delta s = 0$, $\Delta t = 0$ as function of $\Delta x$. (c) $\Delta s = 0$,$\Delta x = 0$ as function of $\Delta t$

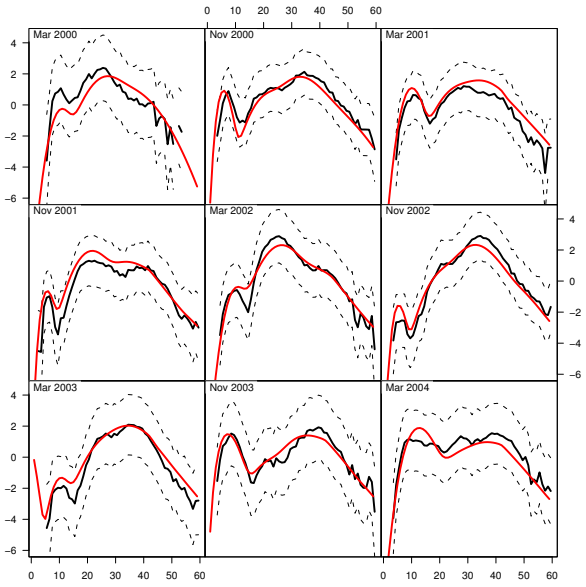



**Fig. 5.** $\hat{\beta}_{size,survey}$ estimated from pure statistical model (black solid line) with 95% confidence limits (broken lines). Fit of first accepted spectra-model (red solid lines).

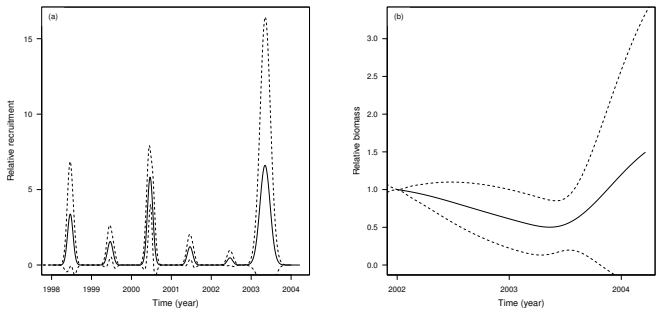



**Fig. 7.** (a) Estimated recruitment $r(t)$ with 95% confidence limits. (b) Estimated relative biomass with 95% confidence limits.

**Table 2.** Estimated parameters and parameter-correlations.

| Description | Parameter | Estimate | Sd | Corr | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Size correlation | $\log(1-\phi_1-\phi_2)$ | -4.33 | 0.07 | 1.00 | | | | | |
| Size correlation | $\phi_1-\phi_2$ | 2.31 | 0.04 | -0.30 | 1.00 | | | | |
| Variance | $\log\sigma^2$ | 1.43 | 0.09 | -0.69 | 0.22 | 1.00 | | | |
| Spatial correlation | $\log b_1$ | -4.24 | 0.18 | 0.11 | -0.06 | -0.57 | 1.00 | | |
| Time correlation | $\log b_2$ | 2.15 | 0.23 | 0.01 | -0.00 | -0.19 | 0.19 | 1.00 | |
| Nugget effect | $\log(\nu^{-1}-1)$ | 1.06 | 0.16 | -0.20 | 0.23 | 0.33 | 0.20 | 0.03 | 1.00 |

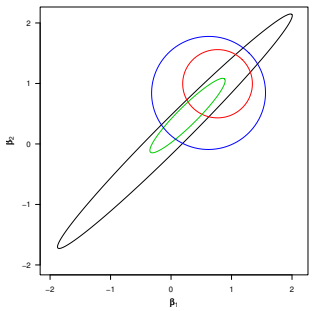



**Fig. 6.** 95%-confidence ellipsis of the pair of $\beta$-parameters corresponding to the neighbor size-classes 20 cm and 21 cm for the spring 2001 survey fitted under different correlation models: No correlation (red). Only space-time correlation (blue). Only size-correlation (green). Both size and space-time correlation (black).

These considerations explains the wide marginal confidence bands around the $\beta$-parameters (Fig. 5). The increased precision of the spectrum-slope is not visible from that figure but would require a different plot (not shown).
After having analysed the main statistical model we can proceed by considering the size-spectrum model hypothesis [8] with parameters $\alpha$ given by [9]. The spectrum model was fitted to data using the quadratic approximation of the LGCP-likelihood around $(\hat{\beta},\hat{\theta})$. The likelihood ratio test of this model against the un-parametrized model was rejected using [11]. Plots of the spectrum model revealed that the rejection was due to wrong position and wideness of the recruitment peak of 2003. This could be seen because the corresponding part of the size-spectrum went outside the confidence bands of the $\beta$-parameters (not shown). By imposing an individual mean-spawning date and wideness for the year in question the size-spectrum model was accepted (p=0.34). The log-size-spectrum of the accepted model lies within the confidence bands of the $\beta$-parameters from the main model (Fig 5).
A further simplification of the model was obtained by testing the asymptotic level of the mortality rate ($F_\infty$) as constant against the alternative of a yearly varying level (p=0.12). A test for no size-selectivity in the total mortality was also accepted (p=0.81).
The resulting spectrum-model has 16 parameters and thus represents a major reduction in complexity compared to the final model of Kristensen et al. (2006) which had 32 parameters.
Recruitment and biomass estimates of the final model are shown in Fig 7. The biomass does not change significantly during the period.
The fact that the LGCP-model accepts the constant catchability hypothesis is an important difference compared to the negative binomial model applied in Kristensen et al. (2006). Tests of constant catchability vs variable catchability is related to the precision of the $\beta$-parameters (Fig. 6). Inclusion of both size and space-time correlation greatly reduces the information about the overall level of the size-spectrum and makes a time variable catchability non-significant. Conversely any of the simpler correlation structures over-estimates the precision of the overall level and therefore rejects the constant catchability hypothesis (Table 3).

**Table 3.** Likelihood ratio test of time-constant catchability hypothesis under different null-models

| Null.model | LR.statistic | p.value |
|---|---|---|
| No correlation at all | 400.27 | 0.00 |
| Size correlation, no space-time-correlation | 33.90 | 0.00 |
| Space-time-correlation, no size-correlation | 185.82 | 0.00 |
| Size, space and time-correlation | 3.95 | 0.86 |

## 5. Discussion

The statistical interpretation of trawl-survey data must account for heterogeneity in order to give valid conclusions about an underlying biological assessment model. A statistical model which correctly describes heterogeneity provides more realistic estimates of the uncertainty associated with biological parameters and predictions.
Complex biological phenomena like Spatio-temporal size-dependent clustering of fish are difficult to explain from first-principles but can relatively easily be modelled on population-level through a correlation function. This geostatistical approach is useful for population dynamical analysis where the main interest is the size- or age distribution of an entire population. Heterogeneity is treated as nuisance but the implications of the heterogeneity are reflected in both biological estimates and in the interpretation of the model.
The present paper follows this idea by combining the log Gaussian Cox process with an existing population model of a single species fish stock. It is thereby investigated how different correlation assumptions affects the final biological conclusions.
The study focused on the random effect of a time changing

large scale intensity landscape and the effect of small-scale size-dependent patchiness. Considering a case study of nine bottom trawl surveys in the Baltic it was found that the correlation structure was necessary to adequately describe the data. Both the effect of a time changing large scale intensity landscape and the effect of small scale size correlation were significant.

Whether the correlations has an impact on the population analysis generally depends on the ranges of the correlations compared to the scale of the study region. For the present case the spatial correlation range constituted 40% of the largest distance between two samples. Similarly the range of the small-scale size correlation constituted more than 100% of the entire length-range. It is thus not surprising that the inclusion of space-time and size correlations turned out to play an important role in the statistical interpretation of the observed size-spectrum.

Generally inclusion of size-correlation increase the precision about the spectrum-slope while decreasing the information about the overall level of the spectrum. The effect of space-time correlation on the precision of the size-spectrum depends on the spatio-temporal coverage of the survey. For the present case the inclusion of space-time correlation decreased the information about the size-spectrum - both the slope and the level.

The implications of a poorly identified level of the size-spectrum for the individual surveys is immediate. Temporal changes of catchability and mortality becomes less significant meaning that statistical tests of time-independence tend to be accepted. This generally leads to biological models with fewer parameters. However, when the number of biological parameters is reduced it generally increases the precision of the remaining parameters. In the end we do not necessarily loose biological information about e.g. biomass but get completely different conclusions.

For the present case a constant catchability hypothesis could be accepted which is a major difference compared with Kristensen et al. (2006) who analysed the same data material with a negative binomial model ignoring the correlations. Most of the biological effects which were significantly time-dependent in the previous study could be tested constant with the new model. The final biological model had only half as many parameters as the final accepted model in Kristensen et al. (2006). The predicted biomass during the period 2002-2004 does not change significantly according to the new model as opposed to the corresponding predictions of Kristensen et al. (2006).

There are many possible explanations of catchability variations (see Harley and Myers 2001) which can roughly be categorized as factors related to the gear and factors related to spatial heterogeneity. The present work attempted to explain apparent temporal catchability variations as an indirect consequence of large-scale spatial heterogeneity and schooling. Other authors (Fryer et al. 2003; Trenkel and Skaug 2005) have modelled between haul variations of catchability directly considering the gear-selection as a stochastic process.

Other statistical distributions for multivariate count data has been applied to bottom trawl surveys in order to capture dependence between size-classes comprising the Dirichlet-multinomial and Gaussian-multinomial models (Hrafnkelsson and Stefansson 2004).

Performing size-based population analysis without accounting for correlations in the data can be dangerous. We have provided a statistical model which consistently deals with correlations caused by various kinds of heterogeneity and shown how to combine it with a length-based population model. The computational requirements of the method is outweighed by a number of advantages. Significance tests for relevant biological complexity are improved compared to previous models and confidence intervals are more reliable. This is because the method automatically accounts for possibly poor spatio-temporal coverage of the survey when calculating confidence intervals. The approach provides an alternative way to model catchability. Instead of treating catchability as a systematic effect to explain catch-variability a similar effect can be obtained through the space, time and size-correlation.

## References

Aitchison, J. and Ho, C. 1989. The multivariate Poisson-log normal distribution. Biometrika, **76**(4):643–653.

Bertalanffy, L. 1938. A quantitative theory of organic growth (Inquiries on growth laws. II). Human Biology, **10**(2):181–213.

Davis, T. 2006a. Direct Methods for Sparse Linear Systems. Society for Industrial Mathematics.

Davis, T. 2006b. User guide for cholmod. Technical report, Tech. rep., University of Florida.

Deriso, R., Maunder, M., and Skalski, J. 2007. Variance estimation in integrated assessment models and its importance for hypothesis testing. Canadian Journal of Fisheries and Aquatic Sciences, **64**(2):187–197.

Fryer, R., Zuur, A., and Graham, N. 2003. Using mixed models to combine smooth size-selection and catch-comparison curves over hauls. Canadian Journal of Fisheries and Aquatic Sciences, **60**(4):448–459.

Harley, S. and Myers, R. 2001. Hierarchical Bayesian models of length-specific catchability of research trawl surveys. Canadian Journal of Fisheries and Aquatic Sciences, **58**(8):1569–1584.

Hrafnkelsson, B. and Stefansson, G. 2004. A model for categorical length data from groundfish surveys. Canadian Journal of Fisheries and Aquatic Sciences, **61**(7):1135–1142.

Jardim, E. and Ribeiro, P. 2007. Geostatistical assessment of sampling designs for Portuguese bottom trawl surveys. Fisheries Research, **85**(3):239–247.

Kristensen, K. 2008. Spatio-temporal modelling of population size-composition with the log gaussian cox process using trawl survey data. Submitted to biometrics.

Kristensen, K., Lewy, P., and Beyer, J. 2006. How to validate a length-based model of single-species fish stock dynamics. Canadian Journal of Fisheries and Aquatic Sciences, **63**(11):2531–2542.

Møller, J., Syversveen, A., and Waagepetersen, R. 1998. Log Gaussian Cox processes. Scand. J. Stat., **25**:451–482.

Petitgas, P. 2001. Geostatistics in fisheries survey design and stock assessment: models, variances and applications. Fish and Fisheries, **2**(3):231–249.

R Development Core Team 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Sibert, J., Hampton, J., Fournier, D., and Bills, P. 1999. An advection-diffusion-reaction model for the estimation of fish movement parameters from tagging data, with application to skipjack tuna (Katsuwonus pelamis). Canadian Journal of Fisheries and Aquatic Sciences, **56**:925–938.

Skaug, H. and Fournier, D. 2006. Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. Computational Statistics and Data Analysis, **51**(2):699–709.

Trenkel, V. and Skaug, H. 2005. Disentangling the effects of capture efficiency and population abundance on catch data using random effects models. ICES Journal of Marine Science: Journal du Conseil, **62**(8):1543.

Wang, Y. and Ellis, N. 2005. Maximum likelihood estimation of mortality and growth with individual variability from multiple length-frequency data. Fish. Bull., **103**:380–391.

## 6. Appendix

### 6.1. Parametrization

The parametric forms of the biological processes are obtained from (Kristensen et al. 2006). The survey size-selection is chosen as

$$\text{sel}(s) = \frac{1}{1+\exp(-\gamma(s-L_{50}^{\text{survey}}))} \quad [12]$$

Total mortality is given by

$$z(s,t) = M_0 + f(s,t) \quad [13]$$

where $M_0$ is an unknown constant and the fishing mortality $f$ is assumed to split into the product of a piecewise constant function of time and a sigmoid function of size

$$f(s,t) = \frac{1}{1+\exp(-\delta(s-L_{50}^{\text{fishery}}))} \sum_{i=1}^{n} F_{\infty}^{(i)} 1_{(\tau_{i-1}<t<\tau_i)}$$

For the distribution of $L_\infty$ $(u)$ we use a normal-distribution with mean $\mu_{L_\infty}$ and standard deviation $\sigma_{L_\infty}$.

$$r(t) = \sum_{y \in Y} R_y \phi_{\mu_y^{recr},\sigma_y^{recr}}(t) \quad [14]$$

where $\phi_{\mu,\sigma}$ is the normal density with mean $\mu$ and standard deviation $\sigma$. The mean recruitment time for cohort $y$ is parameterized as a year $y$ plus a date $\Delta t_y$ i.e. $\mu_y = y + \Delta t_y$.

### 6.2. Stationary AR(2)-process

The AR(2)-process is defined through the recursion

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon_t$$

where $(\epsilon_t)$ are independent with distribution $N(0,\sigma^2)$. When $(\phi_1,\phi_2)$ belongs to the triangular region

$$\{(\phi_1,\phi_2) : \phi_2 > -1,\ \phi_2 < 1+\phi_1,\ \phi_2 < 1-\phi_1\}$$

it is well known that $x_t$ has a stationary initial distribution $\pi(x_1,x_2)$. The correlation function can be found by the so-called Youle-Walker equations:

$$\rho_{size}(0) = 1,\ \ \rho_{size}(1) = \frac{\phi_1}{1-\phi_2}$$

$$\rho_{size}(\Delta s) = \phi_1 \rho_{size}(\Delta s - 1) + \phi_2 \rho_{size}(\Delta s - 2),\ \Delta s \geq 2$$

### 6.3. Implementation

The computational methods of this paper are implemented as R-packages (R Development Core Team 2008) available on request.

# Paper IV

# Modelling the distribution of fish accounting for spatial correlation and overdispersion

Peter Lewy and Kasper Kristensen

P. Lewy[1] and K. Kristensen. National Institute of Aquatic Resources, Technical University of Denmark, Charlottenlund Castle, 2920 Charlottenlund, Denmark

[1]Corresponding author: Peter Lewy. National Institute of Aquatic Resources, Technical University of Denmark, Charlottenlund Castle, 2920 Charlottenlund, Denmark. Tel. +4533963368, Fax +4533963333, email: pl@aqua.dtu.dk
K. Kristensen (email: kkr@aqua.dtu.dk)

**Abstract**

The spatial distribution of cod in the North Sea and the Skagerrak was analysed over a 24 year period using the Log Gaussian Cox Process (LGCP). In contrast to other spatial models of the distribution of fish LGCP avoids roblems with zero observations and includes the spatial correlation between observations. It is therefore possible to predict and interpolate unobserved densities at any location in the area. This is important for obtaining unbiased estimates of stock concentration and other measures depending on the distribution in the entire area. Results show that the spatial correlation and dispersion of cod catches remained unchanged during winter throughout the period in spite of a drastically decline in stock abundance and a movement of the centre of gravity of the distribution towards north east in the same period. For the age groups considered the concentration of the stock was found to be constant or declining in the period. This means that cod does not follow the theory of density-dependent habitat selection as the concentration of the stock does not increase when stock abundance decreases.

**Introduction**

Knowledge of the spatial distribution of fish and the temporal changes are important for the fishery, fishery management and for understanding the mechanisms of fish behaviour. The distribution of cod has been analysed in several studies (Rindorf and Lewy 2006; Perry et al. 2005).These analyses used a single point, the centre of gravity, as an overall measure to describe changes in the spatial distribution. However, if we want to study the spatial distribution of stock abundance in the entire area another type of modelling is required.

Previously fishery scientific survey data have been analysed assuming that observations are independent irrespective of trawl position and distributed according to either extensions of the log normal (Stefánsson 1996) or the negative binomial distributions (O'Neill and Faddy 2003, Kristensen et al. 2006) . Hrafnkelsson and Stefánsson (2004) presented extensions of the multinomial distribution to account for dispersion and correlation in length measurements samples. To avoid the assumptions of independent observations other authors used kriging to account for spatial correlation in the analysis of trawl and acoustic survey data (Stelzenmüller et al 2005; Rivorard et al. 2000). Kriging methods, however, require that data follows a multivariate normal distribution, an assumption which usually is not fulfilled at least not when part of data consists of zero's. The log(catch+constant) transformation is often applied to avoid this problem, a solution which is problematic, because the results heavily depends on the choice of the constant. Here model we instead use a counting model to describe the discrete catch in number observations (including the zero catch observations) and to account for the spatial correlation between catches. The model, the so-called Log-Gaussian Cox Process, LGCP (Kristensen in prep.; Møller et al. 1998, Diggle and Ribeiro Jr. 2007), is also known as the multivariate Poisson-log normal distribution (Aitchison and Ho 1989) and is a mixture of Poisson distributed observations with mean intensities following a multivariate lognormal distribution. The Poisson process can be regarded as the sampling process generated by the fishing process. The spatial correlation is included by assuming correlation between intensities to be a decreasing function of the distance between them.

The focus of Kristensen in prep. was to develop methods for and implement of ML estimation of the parameters in the LGCP, which hitherto has been estimated by MCMC (Møller et al. 2004). Aspects of predictions and interpolation were not included. These aspects are crucial when estimat-

ing total biomass or the biomass in specified areas, a prerequisite to evaluating the effects of spatial closures and temporal changes of stock concentration.

The objective of this paper is to develop ML based methods for predicting the unobserved intensities at any point in space and to enable goodness-of-fit tests.

The use of the model was illustrated by an analysis of the distribution of North Sea and Skagerrak cod in 1983-2006. The temporal change in the dispersion and spatial correlation was examined and the effect of a range of local hydrographical parameters investigated. Contour plots of the spatial distribution of age group 1 were produced by interpolation. The theory of Density Dependant Habitat Selection as formulated by MacCall (1990) was investigated, i.e. if the spatial distribution of a stock contracts/expands when stock abundance decreases/increases. The analysis will be based on the measure of concentration, *D*95 (Swain and Sinclair 1994) calculated from interpolations of intensities.

**Statistical model**

Let $X_i$ be the catch in number from haul $i$ with a known position, let $\lambda_i$ be the unknown, true intensity at the same position, let $a$ and $d$ be dispersion parameters and finally let $b$ be a spatial correlation parameter. Further, let $\boldsymbol{X} = (X_1,...,X_n)^t$ be the vector of $n$ catch samples covering the area, and let $\lambda = (\lambda_1,...,\lambda_n)^t$ be the corresponding true intensities, where $t$ denotes the transposed of a matrix. It is assumed that the duration of the hauls are the same.

The model considered is a compound Poisson distribution where the conditional distribution of the catch, $X_i$, given the intensity, $\lambda_i$, are independent Poisson distributed variables and where $\boldsymbol{\lambda}$ follows a multivariate lognormal distribution:
:

$$X_i \mid \lambda_i \sim Poisson(\lambda_i) \quad \eta = \begin{pmatrix} \eta_1 \\ . \\ \eta_n \end{pmatrix} = \begin{pmatrix} \ln(\lambda_1) \\ . \\ \ln(\lambda_n) \end{pmatrix} \sim N(\mu, \Sigma) \qquad (1)$$

where

$$\mu = \overbrace{(\mu,...\mu)^t}^{n}$$

The variance/covariance matrix $\Sigma$ is defined by

$$\Sigma_{ii} = VAR(\eta_i) = a + d \qquad a \geq 0 \ \wedge \ d \geq 0$$

where $a+d$ is the dispersion of the model.

This implies that $E(\lambda_i) = e^{\mu+(a+d)/2}$, $i$=1,…$n$

The covariance between two intensities is assumed to be a decreasing function of distance between the haul positions such that it approaches zero when distance increases:

$$\Sigma_{ij} = COV(\eta_i, \eta_j) = a * e^{-b*dist(i,j)} + d * I_{dist(i,j)=0} \qquad (2)$$

where *I* is the indicator function and $dist(i,j)$ denotes the distance in kilometres on the surface of an sphere between position *i* and *j*. The relationship between the distance between two points in kilometre and the corresponding longitudes and latitudes (*lon* and *lat*) is:

$$dist(i,j) = dist(lon_i, lat_i, lon_j, lat_j) =$$
$$6378.1 * \text{acos}\left(\sin(lat_i * c) * \sin(lat_j * c) + \cos(lat_i * c) * \cos(lat_j * c) * \cos((lon_i - lon_j) * c)\right)$$

where $c = \pi / 180$.

The correlation between log intensities is

$$COR(\eta_i, \eta_j) = \frac{a}{a+d} e^{-b*dist(i,j)} \quad i \neq j \qquad (3)$$

If *d* is zero the correlation is $e^{-b*dist(i,j)}$ independent of *a*.

The model contains the four parameters $\theta = \{\mu, a, b \text{ and } d\}$ and the unobserved random effects intensities, $\eta$.

Differences in the duration of the hauls have been ignored and are implicitly included in the small scale, nugget effect, see below.

The interpretation of the model is:

1.
The observed numbers caught in a haul given the intensity is assumed to follow a Poisson distribution. This process is interpreted as the fishery sampling process for instance due to variation of the behaviour of the trawl or fish movements.

2.
The intensities in the sea are assumed to follow a multivariate lognormal distribution where the correlation between intensities is a decreasing function of distance between them. The mean and variance of observations in the LGCP is

$$E = E(X_i) = E(\lambda_i) = e^{\mu + (a+d)/2}$$
$$V = V(X_i) = V(X_i \mid \lambda_i) + V(\lambda_i) = E + (e^{a+d} - 1)E^2$$

If the variance of log intensities, $a+d$, is positive the variance of $X$ is greater than the variance of the Poisson process. Hence $a+d$ can be regarded as overdispersion parameters relative to the Poisson process.

The variance-mean relation of LGCP corresponds to that of the negative binomial distribution, for which $V = E + \text{constant} * E^2$.

3.
The covariance defined in equation (2) consists of a sum of the two terms, which respectively can be considered as a large scale and a small scale component of the process. The large scale component include the large scale variance $a$ and the parameter $b$ ($\geq 0$) measuring the strength of the spatial correlation: When $b$ is small the large scale correlation between intensities is high and vice versa. The scaling of the correlation is measured by $\frac{1}{b}$, which is the distance for which the spatial correlation is 0.37 (if $d = 0$). The small scale variance is $d$, which corresponds to the so called "nugget" effect in geostatistics, may for instance be arise due to fish movements. Fig. 1 illustrates the clear large scale variation due to "spatial" correlation for the case where the "nugget" effect is excluded (solid line) and the superimposed small scale variation due the "nugget" effect (dashed line) blurring the large scale effect.

## Predictions of unobserved intensities at positions with observations available

The likelihood function of $X$ of the LGCP expressed as of function of the parameters, $\theta$, is

$$\left.\begin{aligned} &L(\theta) = \int P(X,\theta,\eta)d\eta = \int e^{-l(X,\theta,\eta)}d\eta \\ &\text{where} \\ &l(X,\theta,\eta) = e^{\sum_{i=1}^{n}\eta_i} - \sum_{i=1}^{n} X_i\eta_i + \frac{1}{2}\ln(\det(\Sigma)) + \frac{1}{2}(\eta - \mathrm{M})^t\Sigma^{-1}(\eta - \mu) + \frac{n}{2}\ln(2\pi) + \sum_{i=1}^{n}\ln(X_i!) \end{aligned}\right\} \quad (4)$$

$l(X,\theta,\eta)$ is the negative log likelihood of $(X,\eta)$.

Laplace approximations have been used to calculate $L(\theta)$ in equation (4) for ML estimation of $\hat{\theta}$ and testing hypotheses (Kristensen 2008).

For the positions, where the observations are available, estimates, $\hat{\eta}_\theta(X)$, of log intensities $\eta \mid X$ for given observations $X$ can be obtained by maximizing $l(X,\theta,\eta)$ defined by equation (4) i.e. $\hat{\eta}_\theta(X) = \arg\max_{\eta} l(X,\theta,\eta)$. As indicated the estimate depends on $\theta$ and $X$. As estimate of $\eta$ we use $\hat{\eta}(X) = \hat{\eta}_{\hat{\theta}}(X)$.

Let $l(\eta \mid X) = l(X,\hat{\theta},\eta)$ denote the likelihood of $\eta \mid X$ .The distribution of $\eta \mid X$ is now approximated by the normal distribution with mean $\hat{\eta}(X)$ using a Taylor expansion of $l(\eta \mid X)$:

$$l(\eta \mid X) - l(\hat{\eta}(X)) \cong 0.5 * (\eta - \hat{\eta}(X))^t \left( \frac{\partial^2 l(\eta \mid X)}{\partial \eta_i \partial \eta_j} \right)^{-1}_{\eta=\hat{\eta}(X)} (\eta - \hat{\eta}(X))$$
$$= 0.5 * (\eta - \hat{\eta}(X))^t (\Sigma^{-1} + D_{\hat{\eta}(X)})(\eta - \hat{\eta}(X))$$

where

$$D_{\hat{\eta}(X)} = \begin{pmatrix} e^{\hat{\eta}_1(X)} & & 0 \\ & \cdot & \\ & & \cdot \\ 0 & & e^{\hat{\eta}_n(X)} \end{pmatrix}$$

I.e. the distribution of $\eta \mid X$ is approximated by the quadratic approximation

$$\eta \mid X \sim N(\hat{\eta}(X), (\Sigma^{-1} + D_{\hat{\eta}(X)})^{-1}) \qquad (5)$$

Calculations using "realistic" parameters indicate that this is a good approximation to the true distribution.

Assuming that the approximation holds the estimator $\hat{\eta}(X)$ equals $E(\eta \mid X)$, which is the posterior minimum variance unbiased estimator of $\eta$ for given *X*.

Using that

$$\Sigma = V(\eta) = V(E(\eta \mid X)) + E(V(\eta \mid X)) \cong V(\hat{\eta}(X)) + E((\Sigma^{-1} + D_{\hat{\eta}(X)})^{-1})$$

we find that

$$V(\hat{\eta}(X)) \cong \Sigma - (\Sigma^{-1} + D_{\hat{\eta}(X)})^{-1} \cong \Sigma_{\hat{\theta}} - (\Sigma_{\hat{\theta}}^{-1} + D_{\hat{\eta}(X),\hat{\theta}})^{-1} \qquad (6)$$

## Spatial interpolation

By analogy to the kriging method the LGCP can be used to spatially interpolate the intensity at positions where no observations exist. The best unbiased prediction of any function of the unobserved intensity is the conditional mean given the observations. In the analyses below we assume that the formulas of the conditional means and variances are based on the true value of the parameters. In practice the true values are replaced by the MLE's.

Assume that we want to predict the intensities, $\lambda_{new} = (\lambda_1, ..., \lambda_m)^t = (e^{\eta_1}, ..., e^{\eta_m})^t$, for *m* new positions without observations. First the log intensities $\eta_{new} = \log(\lambda_{new}) = (\eta_1, ..., \eta_m)^t$ are predicted:

According to eq. (1) the combined set of $\eta$ and $\eta_{new}$ are distributed as

$$\begin{pmatrix} \eta \\ \eta_{new} \end{pmatrix} \sim N\left( \begin{pmatrix} \mu \\ \mu_{new} \end{pmatrix}, \begin{pmatrix} \Sigma & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix} \right)$$

where $\mu_{new} = \overbrace{(\mu,...,\mu)}^{m}$ and where $\Sigma_{12}$ are defined similarly as $\Sigma$.

We know that the conditional distribution of $\eta_{new} \mid \eta$ is normal with mean and variance

$$E(\eta_{new} \mid \eta) = \mu_{new} + \Sigma_{21}\Sigma^{-1}(\eta - \mu) \qquad (7)$$

$$V(\eta_{new} \mid \eta) = \Sigma_{22} - \Sigma_{12}\Sigma^{-1}\Sigma_{12} \qquad (8)$$

By analogy with the predictions of $\eta \mid X$ we want to use $E(\eta_{new} \mid X)$ for spatially interpolation:

From the definition of the LGCP the conditional distribution of $X \mid \eta_{new}, \eta$ only depends on the intensities in the points with observations, $\eta$. According to Brémaud (1991) p. 12 this implies that $\eta_{new}$ and $X$ are conditionally independent given $\eta$ and hence $E(\eta_{new} \mid \eta, X) = E(\eta_{new} \mid \eta)$. This implies that

$$\begin{aligned} E(\eta_{new} \mid X) &= E(E(\eta_{new} \mid \eta, X) \mid X) = E(E(\eta_{new} \mid \eta) \mid X) = E(\mathrm{M}_{new} + \Sigma_{21}\Sigma^{-1}(\eta - \mathrm{M}) \mid X) \\ &= \mathrm{M}_{new} + \Sigma_{21}\Sigma^{-1}(E(\eta \mid X) - M) \cong \mathrm{M}_{new} + \Sigma_{21}\Sigma^{-1}(\hat{\eta}(X) - \mathrm{M}) \end{aligned} \qquad (9)$$

$$\begin{aligned} V(\eta_{new} \mid X) &= V(E(\eta_{new} \mid \eta) \mid X) + E(V(\eta_{new} \mid \eta) \mid X) \cong \Sigma_{12}\Sigma^{-1}V(\eta \mid X)\Sigma_{12}\Sigma^{-1} + \Sigma_{22} - \Sigma_{12}\Sigma^{-1}\Sigma_{12} \\ &\cong \Sigma_{12}\Sigma^{-1}(\Sigma^{-1} + D_{\hat{\eta}(X)})\Sigma^{-1}\Sigma_{12} + \Sigma_{22} - \Sigma_{12}\Sigma^{-1}\Sigma_{12} \end{aligned} \qquad (10)$$

Having determined log intensity $\eta_{new}(X)$ the interpolated values of the intensity $\lambda_{new}(X) = E(\lambda_{new} \mid X)$ can be approximated using a Gaussian posterior approximation based on equations (9) and (10):

$$E(\lambda_{new} \mid X) = \exp(E(\eta_{new} \mid X) + diag(V(\eta_{new} \mid X))/2) \qquad (11)$$

We also wish to predict non-linear functions of $\lambda_{new}(X)$ such as the measure of stock concentration, $D95$, defined below. $E(f(\lambda_{new)}) \mid X) = E(f(e^{\eta_{new}}) \mid X)$ and the variance is calculated by simulation by drawing 100 times from the Gaussian posterior approximation based on equations (9) and (10) and calculating the mean and variance of simulated values of $f(e^{\eta_{new}})$.

The spatial interpolation is performed on a regular fine scaled grid. The scale should be chosen sufficiently fine to obtain a good approximation to the continuous random field.

## Other distribution measures

The ability of the LGCP to perform spatial interpolations of the (unobserved) population intensities makes it possible to obtain unbiased estimates of stock characteristics based on intensities in the entire space. A measure of stock concentration is considered: The measure, *D*x, introduced by Swain and Sinclair (1994) is defined as the proportion of the minimum area containing x% of the stock, i.e $D95_{\text{true}}$ (say) is:

$$D95_{true} = \inf_{A \subset E} \left\{ \frac{|A|}{|E|} : \frac{\int_A \lambda(y)dy}{\int_E \lambda(y)dy} = 0.95 \right\}$$

where $E$ indicates the entire area in consideration and $A$ any sub-area of $E$.
If the area is divided into $n$ equally sized sub-areas and $\lambda_i$ represent the intensity in sub-area $i$ then $D95$ can be approximated by

$$D95 = \frac{m + \dfrac{0.95 - z(m)}{z(m+1) - z(m)}}{n}$$

where $z(m) = \dfrac{\sum_{i=1}^{m} \lambda_{(i)}}{\sum_{i=1}^{n} \lambda_i}$, $m \le n$, where $m$ fulfil that $z(m) \le 0.95 \le z(m+1)$ and where $\lambda_{(i)}$ $i$=1,…,$n$ are the intensities sorted in descending order.

*D95* is greater than zero and less than 0.95. $D95$ is conversely proportional the stock concentration, i.e. the concentration of a stock increases when *D95* decreases. *D95* approaches zero when concentration increases and it equals 0.95 if the intensity is constant in the entire space, i.e. when the concentration is minimal.

The validity of the theory of Density Dependant Habitat Selection was investigated by comparing the relation between $D95$ and stock abundance for 1983-2006, for which the abundance drastically was reduced. According to the theory formulated by MacCall (1990) individuals first occupy habitats with the highest suitability, but as realized suitability of these habitats declines due to increasing population density other previously less suitable unoccupied habitat become colonized. Hence the distribution is characterized by spatially equal realized suitability. If the theory holds *D95* should increase when the stock abundance increases.

Calculation of $D95$ has been based on predicted intensities performed on the regular 50 times 50 grid consisting of 808 points as described above. This procedure ensures that an unbiased estimate of $D95$ is obtained, see the Appendix.

## Analysis of residuals and goodness of fit tests

The residuals can be calculated as $X - e^{\hat{\eta}(X)}$. Maximizing the log likelihood $l(X, \theta, \eta)$ (eq. (4)) with respect to $\eta$ shows that $X - e^{\hat{\eta}(X)} = \Sigma^{-1}(\hat{\eta}(X) - \mu)$ and hence the quantity $R = (\hat{\eta}(X) - \mu)$ is linear transformed residuals scaled by log intensity. We prefer to apply these transformed residuals, which expresses the deviation predicted log intensity and the mean. To obtain standardized residuals of *R* the variance of *R* is needed:

$$V(R) = V(\hat{\eta}(X) - \mu) = \Sigma - E((\Sigma^{-1} + D_{\hat{\eta}(X)})^{-1})$$

As the last term in the expression of the variance is not known, we instead use $R^* = R + u$, where $u \sim N(0, (\Sigma^{-1} + D_{\hat{\eta}(X)})^{-1})$ and for which $V(R^*) = \Sigma$ as modified residuals to circumvent this problem. We now assume that $R^* \sim N(0, \Sigma)$ and accordingly $U^* = L^{-1}R^*$ is used as normal, standardized and independent residuals, where $L$ is the lower Choleski triangle of $\Sigma$.

Now assume that we want to examine if residuals are independent of some specified spatial characteristics, for instance the longitude and latitude. We know that $R^*$ is related to the longitude, while $U^*$ is not. However, if $(L^{-1})_{ij}$ is decreasing when the distance between the points *i* and *j* increases, then a specific element $U_i^* = \sum_j (L^{-1})_{ij} R_j^*$ of *U** only depends on the residuals close to the specified observation. This implies that the residuals may be considered as area specific residuals, which has been applied to examine model deviations according to the longitude and latitude.

Goodness of fit tests for validation of the model have been based on the estimated values of log intensity, $\hat{\eta}$, and the MLE of the other parameters, $\hat{\theta}$. Two tests were considered:

$$T_1 = (\hat{\eta}(X) - \hat{\mu})^t V(\hat{\eta}(X))^{-1} (\hat{\eta}(X) - \hat{\mu})$$

where the variance/covariance matrix is determined by equation (6).

The second test is based on the Kolmogorov-Smirnov test quantity and the quantity $U = L_{V(\hat{\eta})}^{-1}(\hat{\eta} - \hat{\mathrm{M}})$

$$T_2 = \max_x |F_U(x) - N(x)|$$

where $L_{V(\hat{\eta})}$ is the lower Choleski triangle of $V(\hat{\eta})$, $F_U$ is the empirical distribution function of U and *N* is the distribution function of the normal distribution:

If log intensity, $\eta(X)$, for given observations, *X*, is normal distributed i.e. $\eta \mid X \sim N(\hat{\eta}, V(\hat{\eta}))$ then $T_1$ and $T_2$ respectively follow the $\chi^2$ and the Kolmogorov distributions. Even that the assumption of normality probably is reasonable implying that the two distributions may be used as test probabilities we instead simulate the exact distributions of $T_1$ and $T_2$:

1. Estimate the parameters in the model and calculate $T_1$ and $T_2$.

2. Simulate new sets of parameters from the normal distribution $N(\hat{\theta}, \hat{\Sigma})$ 100 times using the parameters estimated.
3. Calculate $T_1$ and $T_2$ for each of the 100 repetitions.

The first test should be two-sided and the second one-sided. Hence, the test probabilities $p_1 = P(T_1 > T_{1,obs})$ and $p_2 = 2 * \min(q_2 . 1 - q_2)$ ,where $q_2 = P(T_2 \geq T_{2,obs})$ ,were calculated. The model is accepted if $p$ is greater that 0.05.

The likelihood ratio test was applied to test successive hypotheses regarding the parameters.

## Application

The LGCP was applied to cod catch rates from the International Bottom Trawl Survey (IBTS) in the North Sea and the Skagerrak in February 1983-2006. IBTS is coordinated by the International Council for Exploration of the Sea (ICES) and data is available on www.ices.dk/datacentre/datras/public.asp. The area is confined within 4ºW and 13ºE longitude and 50ºN and 62ºN latitude. The period 1983 and onwards was chosen because the coverage and the survey gear standardization was better compared to previous years. For the 1. quarter survey the annual number of hauls lies between 322 and 534 with a mean of 390. The area contains 186 statistical rectangles (1º longitude by 0.5º latitude), which were covered twice or more. The gear used is a bottom trawl and the haul positions within the rectangles are random selected among trawlable areas. The haul duration is on average 30 minutes, but in 12% of the hauls taken before 1999 the duration was about 1 hour, which may introduce a bias.

The length of the cod caught was recorded and used to determine the age using age length keys. The spatial distribution using LGCP was studied for each of the age groups 1, 2 and 3 years and older.

The hydrographical data, Depth, bottom temperature and salinity by haul, were provided by ICES' hydrographical database. Data for the stock numbers by year and age were obtained from the ICES working group report (ICES 2006).

## Results

The model was used separately for each combination of the age group 1, 2 and $3^+$ and the years 1983-2006 i.e. for 3*24 = 72 combinations.

First the model LGCP has been used to investigate if the position, the depth, the temperature and the salinity can describe the variation of the cpue, i.e. for given age and year it is assumed that

$$\ln(E(X_{age,year,i})) = \alpha + poly(lon_{age,year,i},2) + poly(lat_{age,year,i},2) + poly(depth_{age,year,i},2) + poly(t_{age,year,i},2) + poly(sal_{age,year,i},2) \quad (12)$$

where $i$ denotes sample number for a given age and year, *lon* the longitude, *lat* the latitude, *depth* the bottom depth in meter, $t$ the temperature in Celsius, *sal* the salinity in ppm, *poly*(.,2) a second degree polynomial and $\alpha$ a parameter. The reason why the covariates enter the right hand side of eq. (10) as a second degree polynomial is that this enables the existence of for instance a preferred tem-

perature with a decreasing preference when moving away from the optimum. The assumption of a log linear mean structure was made to ensure that mean cpue remains positive.

MLE's of the parameters in eq. (12) and their confidence intervals obtained from the Hessian matrix were used to test the significance of the parameters. For all years and age groups and parameters the confidence intervals contained zero. One more run with model where the second degree terms were removed gave the same result. Hence for all age groups and years it was concluded that none of the effects associated with the covariates were found to be significant, i.e. log of the expected value, $\mu$, is constant throughout the area independent of any of the explaining covariates. For this model and for all ages and years the four parameters $\mu, a, b$ and $d$ have been estimated.

Regarding the residual analysis the elements of the inverse Choleski, $(L^{-1})_{ij}$, have been plotted against the distance between points for all three age groups and all 24 years. For all 72 plots the functional relation between the inverse Choleski and the distance is very similar. As an example age group 1 in the middle of the period of 1983-2006, 1994 has been selected. The result is given in the upper panel of Fig. 2, which shows that the inverse Choleski elements actually decrease when the distance increases. It appears that outside a circle of 100 kilometres the corresponding residuals, $R^*$ can be neglected indicating that only residuals $U^*$ within the circles are correlated.

The residuals $U^*$ were plotted against longitude and latitude. For none of the residual plots any trend or systematic pattern was found. Plots again for age group 1 in 1994 are shown in the middle and lower panels of Fig. 2.

The validity of the model has been tested using both the goodness-of-fit test statistics $T_1$ and $T_2$. Both tests resulted in that the model was accepted for all age groups and years using a level of significance of 5 percent.

For each age group separately we tested the hypothesis that the parameters $a, b$ and $d$ remain constant over years. The likelihood ratio test was used for that in the following way: Let $l_y(\theta) = -\log(L_y(\theta))$ denote the likelihood function for year $y$, $\hat{\theta}_y$ the MLE of the parameters and $\hat{H}_y = H_{y,\theta=\hat{\theta}_y}$ the estimated Hessian matrix. For each year we approximate the likelihood function with the second order approximation, i.e.

$$l_y(\theta_y) - l_y(\hat{\theta}_y) \cong (\theta_y - \hat{\theta}_y)' \hat{H}_y (\theta_y - \hat{\theta}_y)$$

Using this approximation the simultaneous likelihood function including all years can be approximated by

$$l(\theta) - l(\hat{\theta}) \cong (\theta - \hat{\theta})' \hat{H} (\theta - \hat{\theta}) \qquad (13)$$

where

$$\hat{H} = \begin{pmatrix} \hat{H}_1 & & & \\ & . & & 0 \\ & & . & \\ & 0 & & . \\ & & & \hat{H}_n \end{pmatrix} \qquad (14)$$

and $\theta = (\theta_1,...,\theta_n)^t$

For the likelihood function approximated in equation (13) linear hypotheses of the form $\theta = A\beta$ can be tested by the likelihood ratio test and using that $\hat{\beta} = (A^t \hat{H} A)^{-1} A^t \hat{H} \hat{\theta}$. The homogeneity of the parameters over years i.e $\theta_y = \theta$ for all *y* was tested by setting $A = (\theta,...,\theta)^t$.

For age group 1 *a*, *b* and *d* were accepted to be constant for all years using a likelihood ratio test ($p$=0.29). For age group 2 *a* and *b* were accepted to be constant for all years except for 1999, 2001 and 2005. The analyses of age group 1 and 2 are the most important because main part of the data consists of positive catches. This is in contrast to age $3^+$ for which the zero proportion is in large as it increases from of level of about 40% to about 60%. Hence, the results for these age groups should treated with caution. For age $3^+$ 1983 clearly was an outlier,which was excluded from the analysis. For the remaining years two levels of the parameter *d* appear to divide the years into two groups: 1984, 1988, 1994, 2001 and 2005 for which the nugget effect, *d*, was not significantly different from zero and the remaining 18 years for which *d* is larger than 0.07. For the latter 18 years *a* and *b* were accepted to be constant ($p$=0.86). The results are summarized in table 1.

Table 1 shows that both the characteristic distance, 1/*b* , the large scale *a* and the small scale variation *d*, are decreasing by increasing age indicating that both the spatial correlation and the overdispersion or patchiness declines for increasing age.

Contour plots and *D*95 was calculated based on interpolated values of stock intensity for a regular 50 times 50 grid covering the North Sea and Skagerrak was chosen (confined within 4ºW and 13ºE longitude and 50ºN and 62ºN latitude). This corresponds to areas of about 27 times 24 km. The areas covering land have been removed, which leaves us with a total of 808 positions for which the intensities should be predicted compared to the average of 390 observations available for each of the years 1983-2006. We also tried the finer 70 times 70 grids. The deviations between the two cases with respect to both the mean intensity and the measure of concentration mentioned below were les than 2% indicating that the 50 times 50 grids results in reliable estimates functions of $\lambda_{new}$.

Contour plots are given for age 1 in Fig. 3. The 1-group was until 1997 mainly situated in the southern North Sea and the Skagerrak but has since changed such that a major part is situated in the Skagerrak. It should be noted that this geographical change of distribution is not in contradiction with that the concentration measured using *D*95 is unchanged. This may for instance take place if high density areas geographically change place. Similarly for 2-group the high density area before 2002 was the northern North Sea and the Skagerrak and hereafter mainly the Skagerrak.

**The validity of the theory of density-dependent habitat selection**

Plots of $D95$ and the 95% confidence limits vs. abundance and year are shown for the age groups 1, 2 and $3^+$ (Fig.4). For age 1 linear regression analysis indicates that $D95$ is independent of stock abundance while $D95$ seems to decline with increasing abundance for age 2 and older. This means that even that stock abundance is decreasing drastically during the period the concentration remains unaffected or decreases and accordingly the theory of density-dependent habitat selection for cod in the North Sea in February/March does not hold.

## Discussion

The LGCP applied to analyse the spatial distribution of fishery survey data is a flexible counting model, which was able to describe the spatial distribution of cod in the North Sea and Skagerrak. The model does not assume that observations are independent, but accounts for possible spatial correlation and enables modelling of separate small and large scale variations. Problems with zero catches are avoided due to the discreteness of LGCP. A method for calculating residuals related to latitude and longitude enabling graphical validation of the model has been developed, which makes it possible to examine possible geographical deviations from the model. Finally, two simulated exact tests have been formulated and implemented to perform goodness-of-fit tests.

One of the most important features of the LGCP introduced is the ability to predict and interpolate unobserved intensities at any location in the area independent of the sampling locations. This ability is important because it makes it possible to obtain unbiased estimates of for instance the stock concentration in the area (see the appendix) or the total sum of individuals or biomass. The expected value of the posterior distribution $E(\eta_{new} \mid X)$ is used as basis for interpolation of the spatial distribution of the intensities as it is a minimum variance estimator of $\eta_{new}$ (Diggle and Ribeiro Jr. 2007). Many authors (e.g. Møller et al. 1998) have used MCMC to simulate the posterior mean, which has the advantage that the estimates are unbiased. In the present paper we have instead used a Gaussian approximation to the posterior distribution to estimate posterior means analytically. Simulations indicate that this assumption is reasonable (Kristensen, Submitted). The analytical approach has the advantage that the convergence problems with MCMC for high dimensional data are avoided and the computer time is reduced. The interpolation by sampling from the posterior distribution technique may further be improved using fast Fourier transform and conditioning by kriging (Rue and Held 2005).

The spatial correlation and the large scale variation of the cod distribution did not change in 1983-2006. This is remarkable as the conditions of the stock in the same period drastically changed as cod abundance declined with about 75 % (ICES 2006), centre of gravity of the North Sea component of the stock moved north east about 200 kilometres (Rindorf and Lewy 2006). This indicates that the spatial correlation and variance for cod in the North Sea and Skagerrak seems to be insensitive to major stock changes in the period.

The stability also applies to the concentration of the stock, which is either unchanged over time (age group 1) or declines a bit (age 2 and older). This implies that the theory of Density Dependent Habitat Selection or other density dependent theories do not apply to cod in the North Sea and Skagerrak in wintertime. This result is in contrast to the results of Blanchard et al. (2004) who analysed data from the English Groundfish Survey in the summer (August/September). The conflicting results may be due to differences in the behaviour of cod in the winter and summer or it could be caused by

bias in the estimation of $D95$ using raw or smoothed data especially for small mean catch rates, see the appendix.

From the point of view of fishery management it is crucial that the concentration does not increase with declining abundance. Other things being equal this means that the mean catch rates will not be retained in the commercial fishery when cod abundance declines. If a concentration took place it could lead to an overestimation of the stock size as was the case for cod off Newfoundland (Atkinson et al. 1997, Hutchings 1996).

Analyses of possible relations between local cod occurrence and local hydrographical parameters such as temperature, salinity, depth, latitude and longitude etc. showed that none of the variables affected the cod distribution. Especially, this means that there was no evidence that adult cod locally move to avoid high or low temperature in the winter for which the range of temperature is $-1^{\circ}$ to $9^{\circ}$. This is in agreement with the results of (Rindorf and Lewy 2006) that the centre of gravity for adult fish was not affected of the average temperature and wind.

The effect of the spatial distribution of the fishery on distribution of the stock is not included in the analyses because of lack of data. If by-catch and discard of the 1-group is limited the effect is of minor importance as the fishing mortality rate for trawl and gill net fishery in relation to the total mortality is small (the proportion is about 0.35 in the period). For the two year old fish and older the effect may be important as the proportion is greater than .6 in the period (ICES 2006).

The interpolated intensities indicate that the 1-group shifted from mainly to be located in the southern North Sea and the Skagerrak to mainly to be situated in the Skagerrak only. This indication of temporal correlation would be valuable to incorporate into the model. If such a model with positive temporal correlation was accepted it would enable annual or seasonal predictions of the spatial distribution of fish stocks.

In conclusion, LGCP is a flexible model of the spatial distribution of fish accounting for spatial correlation between densities and avoiding problems with zero observations. It is therefore possible to interpolate the densities at any location in the area, which for instance could be used in connections with evaluation of the effects of closed areas. The model can be used to test the significance of relations between fish occurrence and hydrographical or climatic factors.

**References**

Aitchinson J., and Brown, J.A.C. 1976. The lognormal distribution. Cambridge University Press. Cambridge.

Aitchinson, J., and Ho, C.H. 1989. The multivariate Poisson-log normal distribution. Biometrika **76**: 643-653.

Atkinson, D.B., Rose, G.A., Murphy, E.F., and Bishop, C.A. 1997. Distribution changes and abundance of northern cod (*Gadus morhua*), 1981-1993. Can. J. Fish. Aquat. Sci. **54**(Suppl. 1): 132-138.

Blanchard, J.L., Mills, C., Jennings, S., Fox, C.J. Rackham, B.D., Eastwood, P.D., and O'Brien, C.M. 2004. Distribution-abundance relationships for North Sea Atlantic cod (*Gadus morhua*): observation versus theory. Can. J. Fish. Aquat. Sci. **62**: 2001-2009.

Brémaud, P. 1999. Markov chains, Gibbs fields, Monte Carlo simulation, and queues. Springer, New York.

Diggle, P.J., and Ribeiro Jr., P.J. 2007. Model-based Geostatistics. Springer.

Hrafnkelsson, B., and Stefánsson, G. 2004. A model for categorical length data from groundfish surveys. Can. J. Fish. Aquat. Sci. **61**: 1135-1142.

Hutchings, J.A. 1996. Spatial and temporal variation in the density of northern cod and a review of hypotheses for the stock's collapse. Can. J. Fish. Aquat. Sci. **53**: 943-962.

ICES. 2006. Report of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak. ICES ACFM:35.

Kristensen, K., Lewy, P., and Beyer, J.E. 2006. How to validate a length-based model of single species fish stock dynamics. Can. J. Fish. Aquat. Sci. **63**: 2531-2542.

Kristensen, K. Spatio-temporal modelling of population size-composition with the log-daussian cox process using trawl survey data. Submitted**??**.

MacCall, A.D. 1990. Dynamic geography of marine fish populations. Washington Sea Grant Program, Seattle, Wash.

Møller, J., Syversveen, A., and Waagepetersen, R. 1998. Log Gaussian Cox Processes. Scand. J. Statist. **25**:451-482.

Murawski, S. A., and Finn, J. T. 1988. Biological bases for mixed-species fisheries: species co-distribution in relation to environmental and biotic variables. Can. J. Fish. Aquat. Sci. **45**: 1720-1735.

Rindorf, A., and Lewy, P. 2006. Warm, windy winters drive cod north and homing of spawners keeps them there. J. Applied Ecol. **43**:445-453.

Rue, H., and Held, L. 2005. Gaussian Markov Random Fields. Chapman & Hall/CRC Boca Raton.

Perry, A.L., Low, P.J., Ellis, J.R., and Reynolds, J.D. 2005. Climate Change and Distribution Shifts in Marine Fishes. Science **308**: 1912:1915.

Rivoirard, J., Simmonds, J., Foote, K.G., Fernandes, P., and Bez, N. 2000. Geostatistcs for Estimating Fish Abundance. Blackwell Science. Oxford.

Schrum, C., St. John, M., and Alekseeva, I. 2006. ECOSMO, a coupled ecosystem model of the North Sea and Baltic Sea: Part II. Spatial-seasonal characteristics in the North Sea as revealed by EOF analysis. Journal of Marine Systems, v. 61, iss. 1-2, p. 100-113.

Stefánsson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES J. Mar. Sci. 53: 577–588.

Stelzenmüller, V., Ehrich, S., and Zauke, G-P. 2005. Effects of survey scale and water depth on the assessment of spatial distribution patterns of selected fish in the northern North Sea showing different levels of aggregation. Mar. Biol. Res. 1: 375-387.

Swain, D.P., and Sinclair, A.F. 1994. Fish distribution and catchability: what is the appropriate measure of distribution? Can. J. Fish. Aquat. Sci. **51**: 1046-1054.

**Appendix**

When calculating the index related to stock concentration, *D*95, one has to ensure that the estimate is non-biased. The biasness of *D*95 will be examined here assuming that data follow the LGCP and that the estimation of *D*95 is based on the estimated LGCP parameters and interpolation onto a 50 times 50 grid. It will further be shown that using the raw or smoothed observations as basis for estimating *D*95 may result in biased estimates for small values of the mean catch rate. Finally, it will be demonstrated that *D*95 is closely related to the dispersion and the spatial correlation of data. The simulation experiments are performed in the following way:

A. Estimation of *D*95

1. *D*95 estimated based on LGCP predictions

An area confined by longitudes $0^o$ to $10^o$ and by latitudes $50^o$ to $60^o$ has been considered for which the maximum distance between the corners is 1280 km. For a regular 51*51( = 2601) grid with longitudes $(0^o,0.2^o,....,10^o)$ and latitudes $(50^o,50.2^o,.....,50^o)$ one realization of the intensities, $\lambda = (\lambda_1,...,\lambda_{2601})^t$, for the 2601 gridpoints has been simulated assuming that they follow a LGCP with known parameters $\mu$ (the common log intensity), *a* (the overdispersion) and *b* (the spatial correlation parameter). A nugget effect is not included. The simulations are performed by first calculating the distances between the 2601 points and – based on that – the variance/covariance matrix, $\Sigma$, using the known parameters *a* and *b* and equation (2). Then the 2601 log intensities, $\eta = (\eta_1,...,\eta_{2601})^t$, are simulated by randomly drawing from the multivariate normal distribution, $N(M,\Sigma)$

where $\mathrm{M}^t = \overbrace{(\log(\mu) - a/2,...\log(\mu) - a/2)}^{2601}$ (A1)

The intensities, $\lambda$, are then $\lambda = e^{\eta} = (e^{\eta_1},...,e^{\eta_{2601}})^t$ for which $E(\lambda_i) = \mu$, *i*=1,…2601. Based on the 2601 values of $\lambda$ *D*95 has been calculated. For the selected values of the parameters it has been shown that a 51*51 grid is sufficient to obtain an estimate of the true *D*95, for which the error is less than 0.01. Hence we consider this estimate, $D95_{true}$, as the true *D*95 for the realized distribution of intensities.

We now simulate the catches $X = (X_1,...,X_{121})^t$ on the 11*11 = 121 grid with longitudes $(0^o,1^o,....,10^o)$ and latitudes $(50^o,51^o,....,60^o)$ which is a regular subset of the 51*51 grid. This grid approximately corresponds to that one haul is taken in 60 times 60 nautical miles statistical square, which is a rough grid with a poor covering of the area considered. The catches are simulated by randomly drawing from the independent Poisson distributions with means equal to the corresponding subset, $\lambda_{11*11}$, of $\lambda$.From the simulated catches, *X*, the estimates of the parameters, $\hat{\mu}$, $\hat{a}$ and $\hat{b}$ have been obtained by ML and based on that the predictions on the 51*51 grid of the log intensities $\hat{\eta}_{predicted} = (\hat{\eta}_1,...,\hat{\eta}_{2601})$ and the variance have been calculated using equations (9) and (10). In principle an estimate of $\hat{\lambda}_{predicted} = \exp(\hat{\eta}_{predicted} + \hat{a}/2)$ could be obtained, but as this may be seriously biased (Aitchinson and Brown 1976) we instead simulate an unbiased estimate by 1. Drawing from the multivariate normal distribution $\hat{\eta}_{sim} = N(\hat{\eta}_{predicted}, VAR(\hat{\eta}_{predicted}))$ 2. Calculating

$\hat{D}95_{sim} = D95(\exp(\hat{\eta}_{sim}))$ 3. Repeating 1. and 2. 1000 times and calculating the $\hat{D}95_{LGC}$ =the mean of $\hat{D}95_{sim}$. The possible bias of $\hat{D}95_{LGC}$ has been calculated as $\hat{D}95_{LGC} / D95_{true} - 1$.

2. *D*95 estimates based on the observations

As some authors use the observations or smoothed value of these as basis for estimating *D*95 (Swain and Sinclair 1994, Atkinson et al. 1997, Blanchard et al. 2004) we also examine the possible bias by calculating $D95_{\text{observations}}(X)$ based on the (raw) 121 observations and two alternatives based on smoothed values of. The first estimate, $D95_{81}(X)$, based on smoothed observations is obtained by dividing the intervals [0,10] and [50,60] defining the area considered into 9 intervals and calculating the mean of the observations in each of the 9*9 = 81 rectangles defined. Correspondingly, $D95_{25}(X)$ is obtained by dividing into 5 intervals.

B. The relationship between *D*95 and the spatial correlation and the dispersion

To examine the above relationship $D95_{\text{true}}$ has been calculated for a range values of 1/*b* with fixed dispersion *a* and vice versa.

**Results**

A. Estimation of *D*95

To examine the effect of varying mean catch rate the following sets of simulations have been performed for fixed values of *a* and *b* for the following values of the catch rate:

$a = 1.5 \;\; b = 0.005$ catch rates = $(0.25, 0.4, 0.50, 0.75, 1, 1.5, 2, 3, 5, 7, 8, 10)$ i.e. $\mu = \ln(\text{catch rate})$

For *b* = 0.005 the characteristic spatial correlation distance 1/*b* = 200 km which is the distance for which the spatial correlation equals 0.37. For the case considered the proportion of the points for which the distance is less than 200 km is about 10% indicating that observations are available for estimating the parameters in the model. The coefficient of variation of the intensities $\lambda$ is $\sqrt{1.5} = 1.22$.

The results of the simulations are given in Fig. A1 showing the relationship between the relative bias of estimates of *D*95 and the mean catch rate. The Figure shows that in general $\hat{D}95_{LGC}$ is the least biased estimator of *D*95 and that the bias is less than 5% for mean catch rates larger than 1. For mean catch rates less than 1 the relative bias is less than 15%. The smoothed estimate, $D95_{81}$ is the second best estimator for which the relative bias is less than 10% for mean catch rates larger than 2. For mean catch rates less than 2, however, the relative bias is tremendous, up to about –60%. The bias of $D95_{25}$ in general seems to be positive, up to 20%. For small values of mean catch rates the bias is still limited, below10%. $D95_{\text{observations}}$ is negatively biased and especially for small values of the mean catch rate the bias huge (up to - 60%).

We conclude that the LGCP estimator, $\hat{D}95_{LGCP}$, is the best estimator. Smoothing of the observations may lead to satisfactory *D*95 estimates for mean catch rates larger than 1 or 2. Raw observations should not be used for estimation of *D*95.

B. The relationship between $D95$ and the spatial correlation and the dispersion

The results are shown in Fig, A2. The upper panel shows the relationship between $D95$ and $1/b$ for fixed values of the mean catch rate of 10 and of $a = 1.5$, while the lower panel shows relationship between $D95$ and $a$ for a mean catch rate of 10 and $b = 0.02$. The figure shows that $D95$ depends both on the spatial correlation the dispersion and of the log intensities. Hence, changes in the concentration of a stock (1-$D95$/0.95) may be caused either by changes in the dispersion or the spatial correlation or changes in both stock characteristics. The quantity $1 - D95/0.95$ is measure of the concentration which lies between zero and 1.

**Figure captions**

Fig. 1. Simulation of a LGCP without a "nugget", small scale variation effect (solid line) and the same process including a positive "nugget" effect (dashed line).

Fig. 2. Plot of the relationship between the elements of the inverse, lower Choleski triangle, $L^{-1}$, and the distance between corresponding points (upper panel) and residuals plotted against the longitude (middle panel) and the latitude (lower panel) for age group 1 in 1994. See text.

Fig. 3. Contour plots for 1 year old cod in the North Sea and Skagerrak 1983-2006 based on interpolation onto a 50 times 50 grid.

Fig.4. Minimum area occupied by 95% of the stock, $D95$, by age plotted against stock number for cod in the North Sea and Skagerrak 1983-2006 (solid line) and 95% confidence limits (dashed lines). The straight lines indicate linear regression lines.

Fig. A1. Relative bias of estimated $D95$ versus mean catch rate. Thick solid line: Estimates based on predictions using the LGCP. Solid line: Estimates based on raw catch observations. Dashed and dotted lines: Estimates based on smoothed catch observations. See text in the appendix.

Fig. A2. The relationship between $D95$ and $1/b$ (upper panel) and the variance of log intensity $a$ (lower panel).

**Tables**

Table 1. Estimated parameters and the 95% confidence limits (L95% and U95%) by age

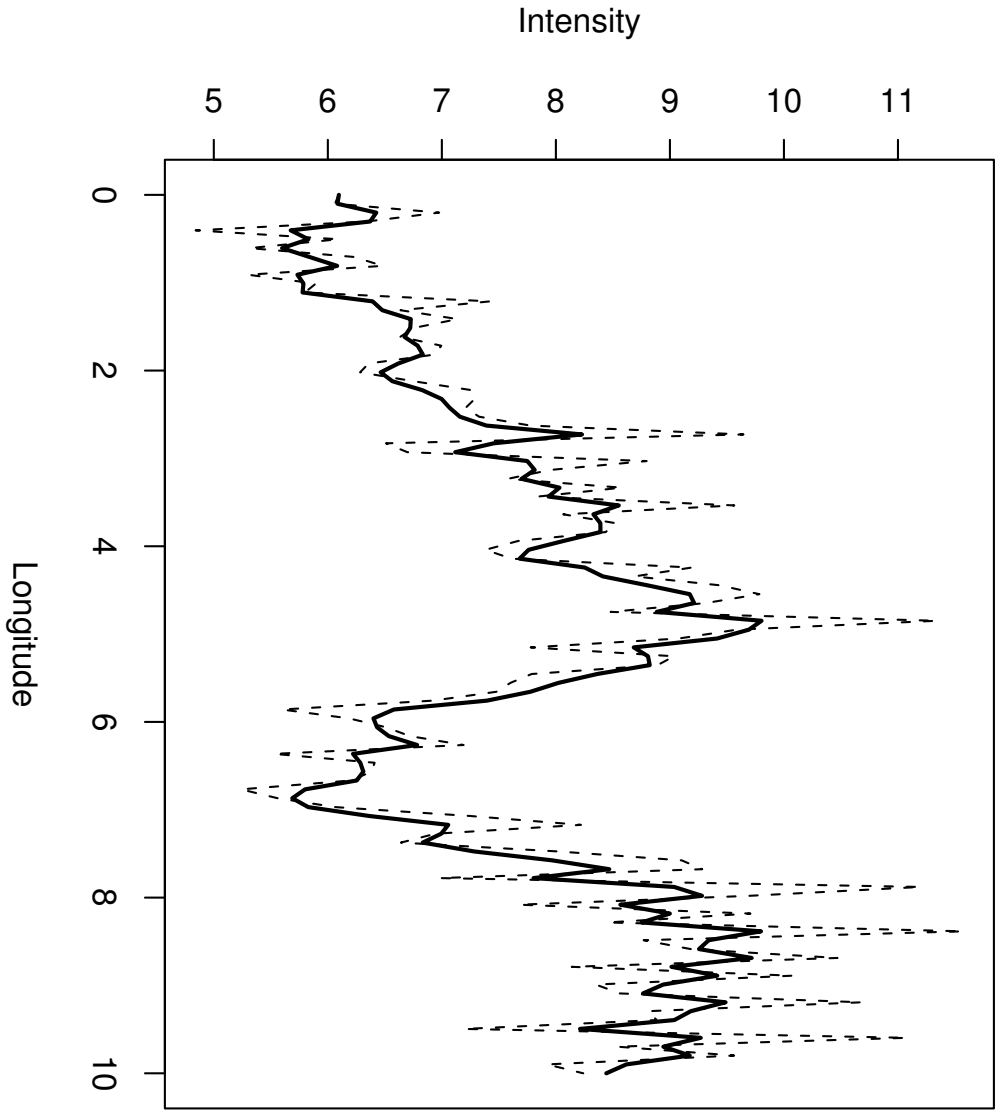
| | Age 1 | | | [2]Age 2 | | | [3]Age $3^+$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | L95% | Mean | U95% | L95% | Mean | U95% | L95% | Mean | U95% |
| [4]1/*b* | 179.1 | 242.0 | 327.0 | 66.2 | 87.2 | 114.5 | 44.8 | 58.9 | 77.3 |
| *a* | 4.08 | 4.98 | 6.07 | 2.11 | 2.41 | 2.75 | 0.98 | 1.12 | 1.27 |
| *d* | 1.26 | 1.42 | 1.58 | | 1.02[1] | | | 0.71[1] | |

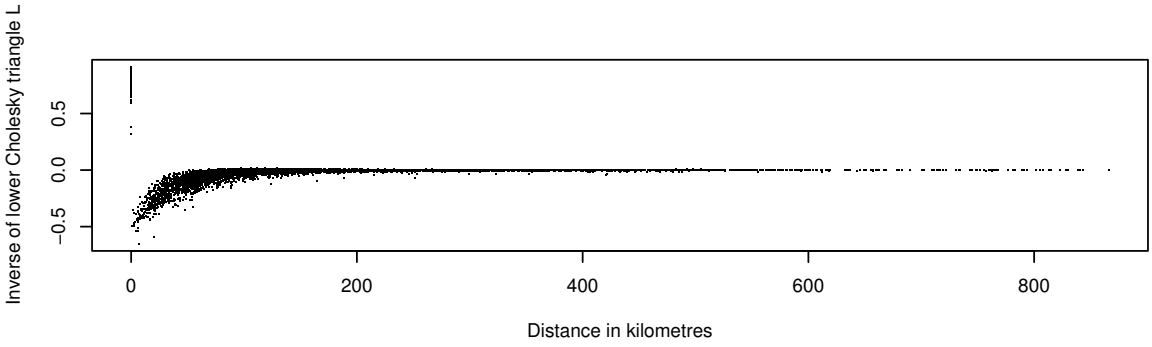
[1] $\exp(\overline{\log(d)})$

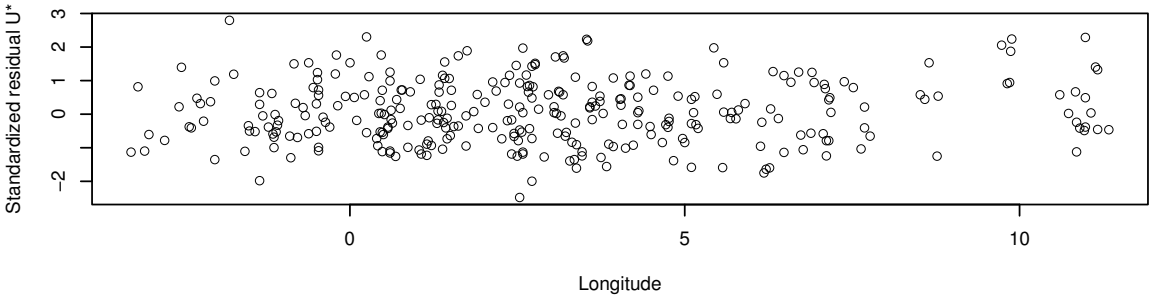
[2] 1999, 2001, 2005 excluded

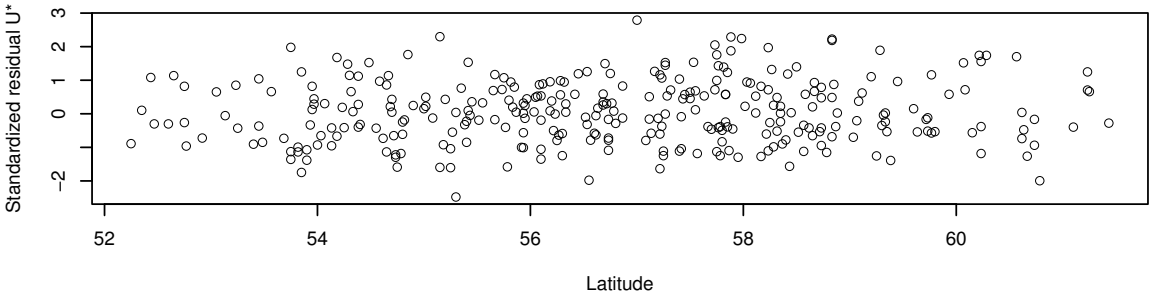
[3] 1983, 1984, 1988, 1994, 2001, 2005 excluded

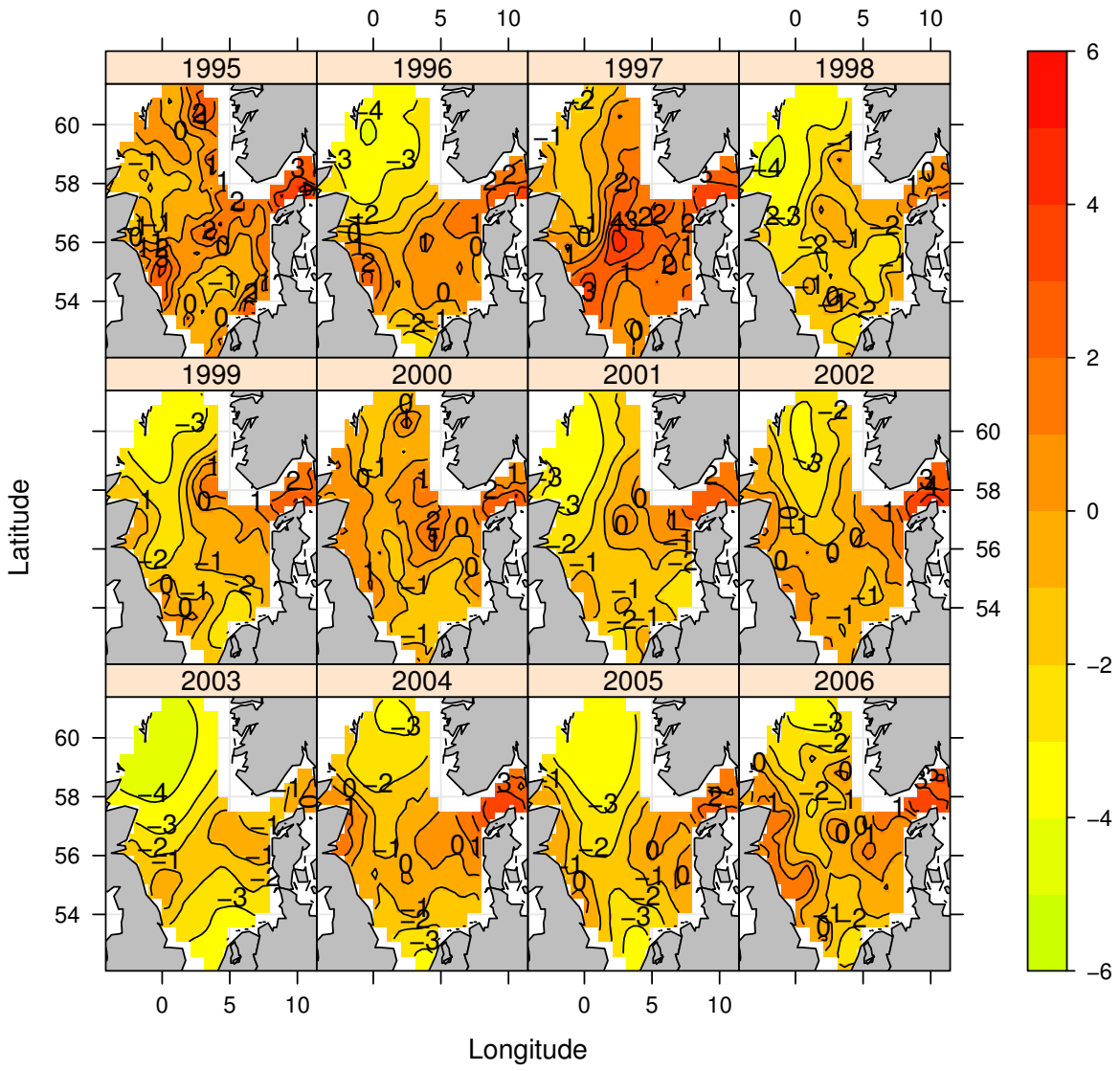
[4] 1/*b*, the characteristic distance, is the distance in kilometre for which the correlation between log intensities is 0.37.

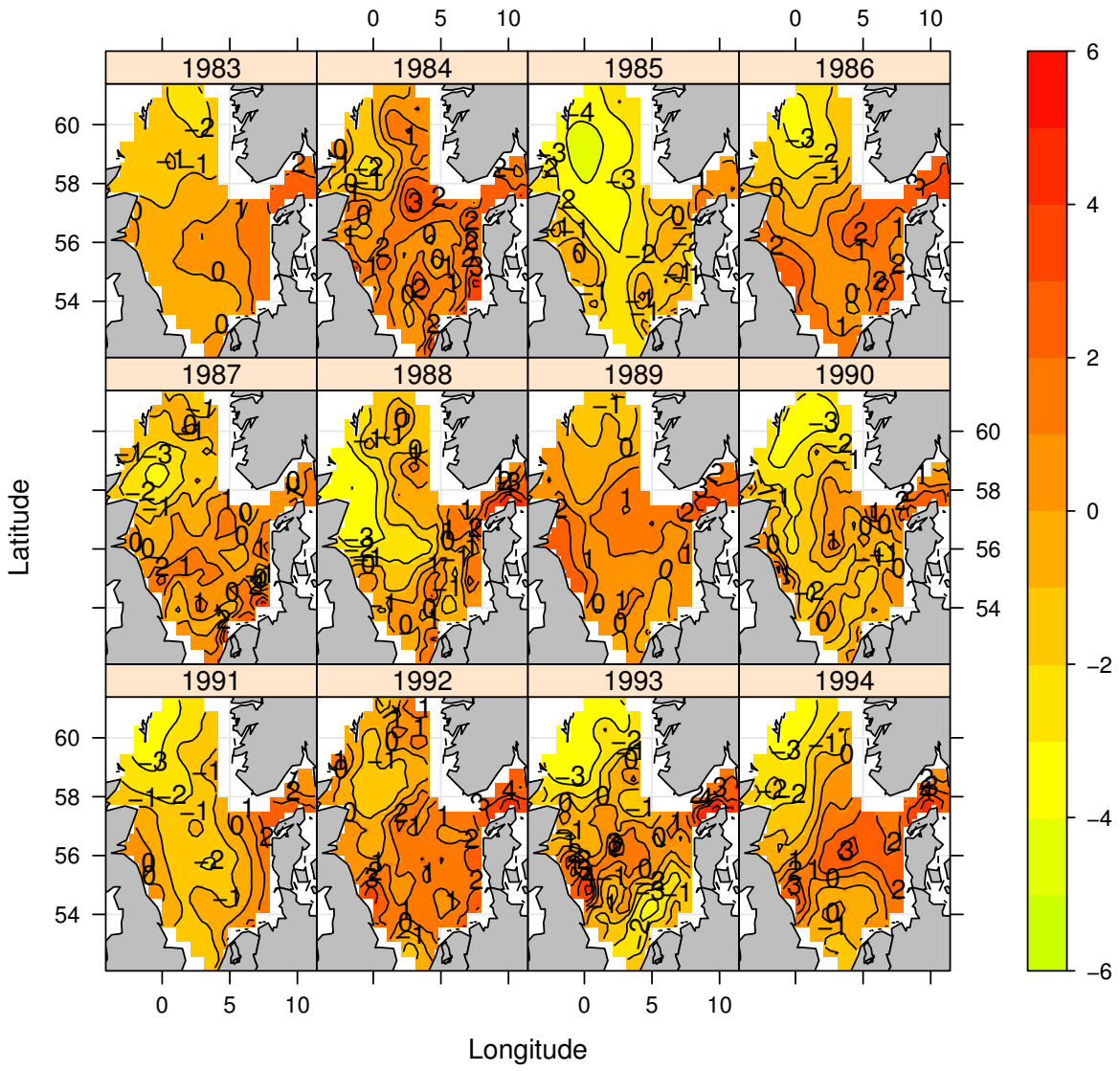

Intensity
5
6
7
8
9
10
11
Longitude
0
2
4
6
8
10

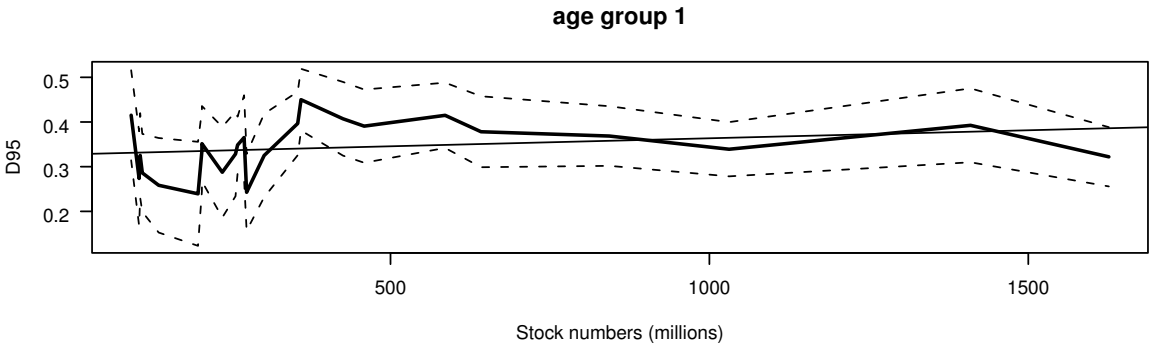

Inverse of lower Cholesky triangle L
−0.5
0.0
0.5
0
200
400
600
800
Distance in kilometres

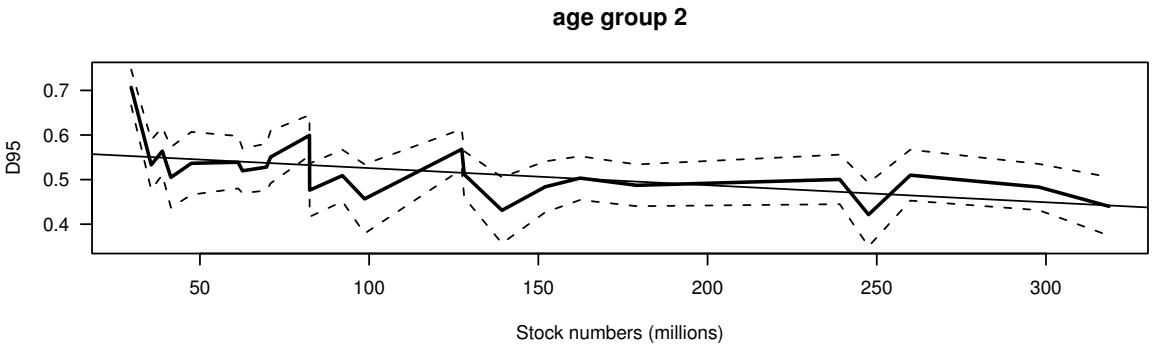

Standardized residual U*
−2
0
1
2
3
0
5
10
Longitude

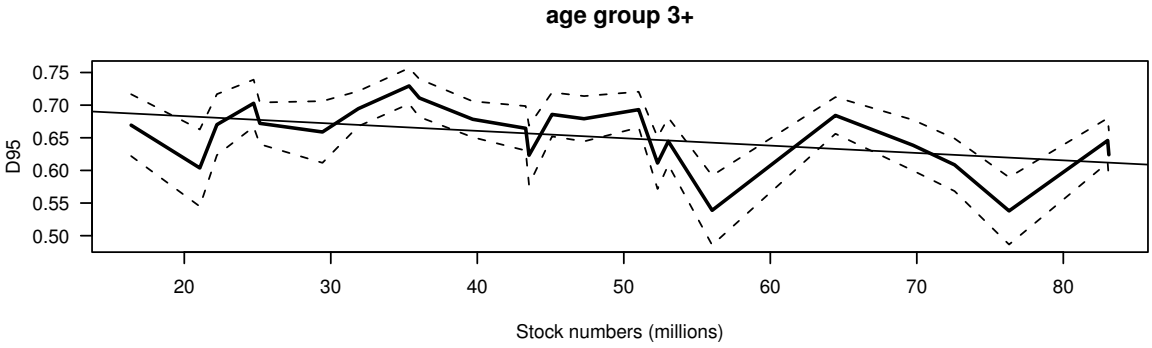

Standardized residual U*
−2
0
1
2
3
52
54
56
58
60
Latitude

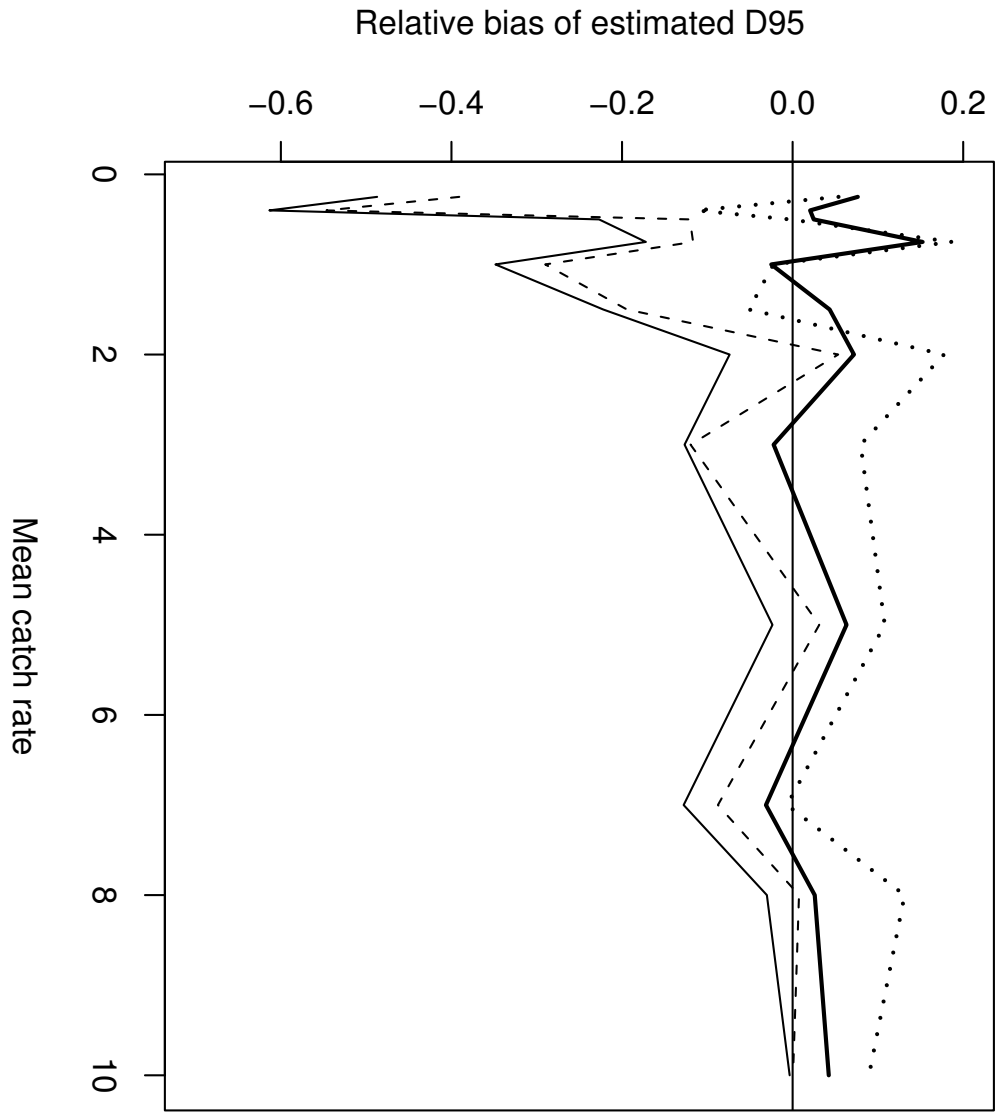

1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
Latitude
Longitude
60
58
56
54
0
5
10
6
4
2
0
−2
−4
−6

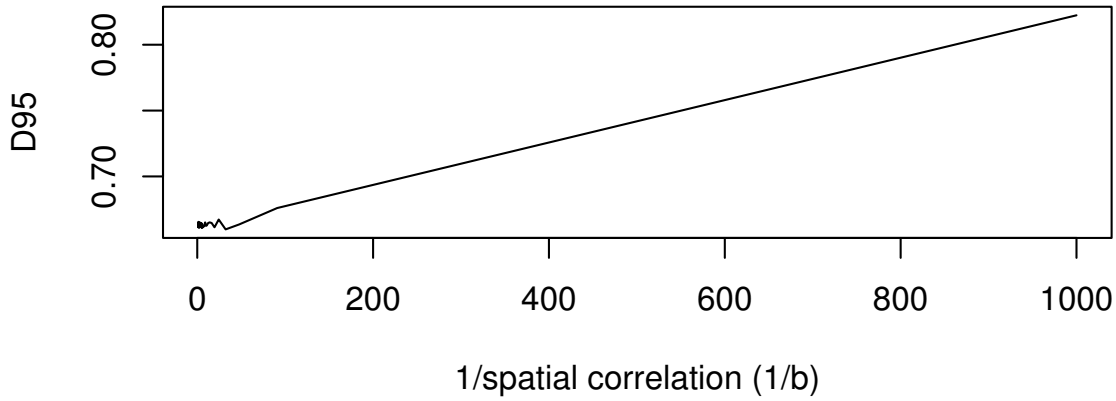

1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
Latitude
Longitude
60
58
56
54
0
5
10
6
4
2
0
−2
−4
−6

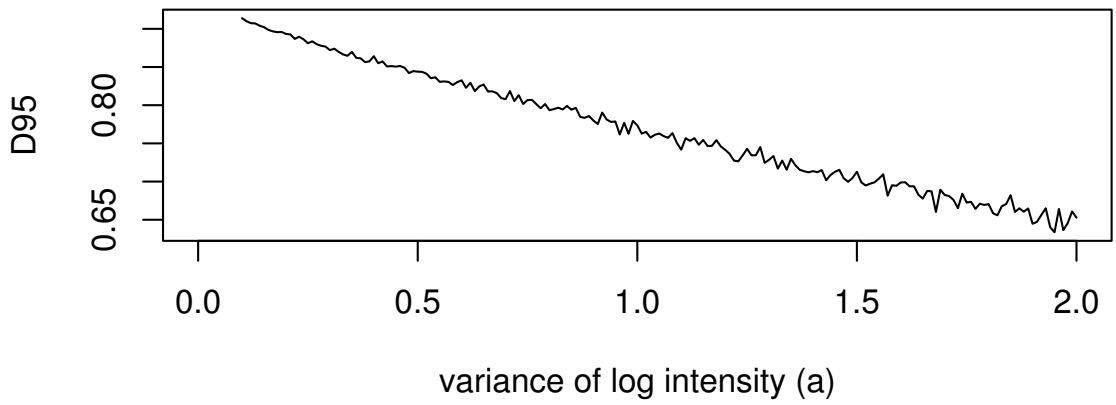

age group 1
D95
0.5
0.4
0.3
0.2
500
1000
1500
Stock numbers (millions)


age group 2
D95
0.7
0.6
0.5
0.4
50
100
150
200
250
300
Stock numbers (millions)


age group 3+
D95
0.75
0.70
0.65
0.60
0.55
0.50
20
30
40
50
60
70
80
Stock numbers (millions)


Relative bias of estimated D95
−0.6
−0.4
−0.2
0.0
0.2
Mean catch rate
0
2
4
6
8
10


D95
0.70
0.80
0
200
400
600
800
1000
1/spatial correlation (1/b)


D95
0.65
0.80
0.0
0.5
1.0
1.5
2.0
variance of log intensity (a)